



This is a repository copy of *MHC Hammer reveals genetic and non-genetic HLA disruption in cancer evolution*.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/219031/>

Version: Published Version

Article:

Puttick, C. orcid.org/0009-0006-3022-8737, Jones, T.P. orcid.org/0000-0003-2748-0780, Leung, M.M. orcid.org/0000-0001-9555-1080 et al. (279 more authors) (2024) MHC Hammer reveals genetic and non-genetic HLA disruption in cancer evolution. *Nature Genetics*, 56. pp. 2121-2131. ISSN 1061-4036

<https://doi.org/10.1038/s41588-024-01883-8>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>


MHC Hammer reveals genetic and non-genetic HLA disruption in cancer evolution

Received: 16 December 2022

Accepted: 25 July 2024

Published online: 2 October 2024

 Check for updates

Clare Puttick ^{1,2,3}, Thomas P. Jones ^{2,3}, Michelle M. Leung ^{1,2,3}, Felipe Galvez-Cancino ^{3,4,114}, Jiali Liu ⁴, Manuel Varas-Godoy ^{5,6}, Andrew Rowan¹, Oriol Pich ^{1,3}, Carlos Martinez-Ruiz ^{1,2,3}, Robert Bentham^{2,3}, Krijn K. Dijkstra^{1,3,7,8}, James R. M. Black ^{1,2,3}, Rachel Rosenthal¹, Nnennaya Kanu ³, Kevin Litchfield ⁹, Roberto Salgado ^{10,11}, David A. Moore ^{1,3,12}, Peter Van Loo ^{13,14,15}, Mariam Jamal-Hanjani^{3,16,17}, Sergio A. Quezada ^{3,4}, TRACERx Consortium*, Charles Swanton ^{1,2,17}  & Nicholas McGranahan ^{2,3} 

Disruption of the class I human leukocyte antigen (HLA) molecules has important implications for immune evasion and tumor evolution. We developed major histocompatibility complex loss of heterozygosity (LOH), allele-specific mutation and measurement of expression and repression (MHC Hammer). We identified extensive variability in HLA allelic expression and pervasive HLA alternative splicing in normal lung and breast tissue. In lung TRACERx and lung and breast TCGA cohorts, 61% of lung adenocarcinoma (LUAD), 76% of lung squamous cell carcinoma (LUSC) and 35% of estrogen receptor-positive (ER+) cancers harbored class I HLA transcriptional repression, while HLA tumor-enriched alternative splicing occurred in 31%, 11% and 15% of LUAD, LUSC and ER+ cancers. Consistent with the importance of HLA dysfunction in tumor evolution, in LUADs, HLA LOH was associated with metastasis and LUAD primary tumor regions seeding a metastasis had a lower effective neoantigen burden than non-seeding regions. These data highlight the extent and importance of HLA transcriptomic disruption, including repression and alternative splicing in cancer evolution.

Emerging data have highlighted the importance of considering cancer evolution in the context of a predatory immune microenvironment^{1–3}. Key mediators of the cytotoxic T cell response in cancer are neoantigens, cancer-cell-specific alterations resulting in mutant peptides capable of eliciting a T cell-mediated, human leukocyte antigen (HLA)-restricted immune response. A mutation can only result in a neoantigen if the associated mutant peptide is presented on HLA molecules to the T cell receptor. Therefore, disruption of HLA molecules has important implications for immune evasion.

Disruption to antigen-presenting machinery occurs across many cancer types^{4–7}. Our previous work has revealed that HLA loss of heterozygosity (LOH), whereby one allele is somatically lost, occurs in 40% of non-small cell lung cancer (NSCLC) primary tumors⁷. A pan-cancer study has suggested that transcriptomic downregulation of HLA genes occurs frequently⁴. However, subtle transcriptomic alterations, such as alternative splicing events and allele-specific repression, have been poorly studied in cancer. Alternative splicing, which has been reported in non-cancer tissue and cancer cell lines, can result in a non-functional

A full list of affiliations appears at the end of the paper. *A list of authors and their affiliations appears at the end of the paper.

 e-mail: charles.swanton@crick.ac.uk; nicholas.mcgranahan.10@ucl.ac.uk

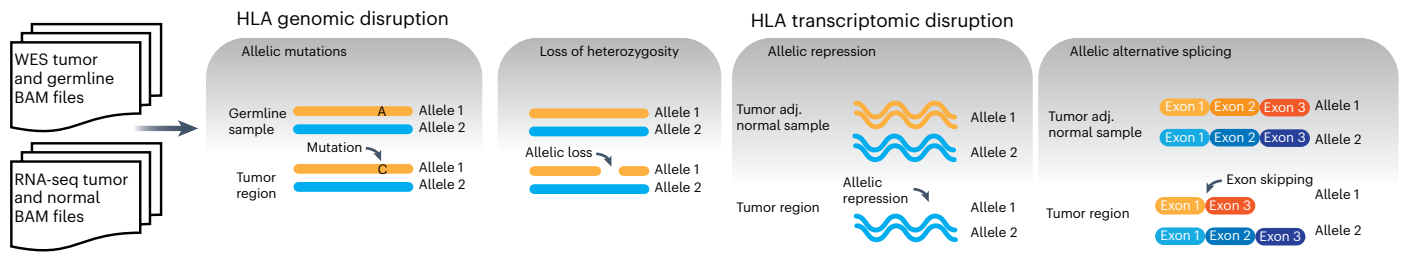


Fig. 1 | MHC Hammer: a tool to evaluate HLA DNA and RNA disruption. MHC Hammer assesses allelic mutations, LOH, allelic repression and allelic alternative splicing in the class I HLA genes from WES and RNA-seq data. tumor adj, tumor adjacent; WES, whole-exome sequencing; RNA-seq, RNA sequencing.

HLA molecule or, in the case of exon 5 skipping, soluble isoforms of the HLA molecule^{8–13}. Furthermore, understanding HLA expression in tumor-adjacent normal tissue is of critical importance when attributing any change in HLA expression as a tumor-specific phenomenon.

Here we present major histocompatibility complex loss of heterozygosity, allele-specific mutation and measurement of expression and repression (MHC Hammer), a computational toolkit to accurately determine allele-specific mutations, LOH, allelic expression, allelic repression and alternative splicing of the class I HLA genes. We use MHC Hammer to investigate HLA expression in normal tissue and to evaluate genomic and transcriptomic disruption in tumor evolution in multiple cohorts, including 421 patients with NSCLC in the multiregional TRACERx421 dataset^{14–16} (Extended Data Fig. 1), 945 patients with NSCLC and 972 patients with breast cancer in the Cancer Genome Atlas (TCGA) dataset^{17–19} (Extended Data Fig. 2) and 489 normal lung and 397 normal breast samples from the Genotype-Tissue Expression (GTEx) dataset²⁰ (Extended Data Fig. 3).

Results

A pipeline to evaluate HLA disruption

To evaluate the extent of genomic and transcriptomic HLA disruption, we developed MHC Hammer, advancing our LOHHLA algorithm⁷ (Fig. 1). The tool has the following four major components: (1) identifying allele-specific HLA somatic mutations, (2) calculating HLA LOH, (3) evaluating HLA allele-specific repression and (4) identifying allele-specific HLA alternative splicing. MHC Hammer is provided as a Nextflow pipeline (<https://github.com/McGranahanLab/mhc-hammer>) (Methods; Supplementary Note and Supplementary Fig. 1).

HLA allele-specific expression in normal tissue

We first evaluated HLA allelic expression and alternative splicing in normal lung and breast tissue using data from the GTEx project²⁰. This dataset includes 489 lung and 397 breast tissue samples from 645 healthy individuals, of which 241 have both lung and breast tissue samples available (Methods; Extended Data Fig. 3).

We found that in both normal lung and breast tissue, HLA-B had the highest median expression (lung, 440.3 and breast, 227.4 reads per kilobase million (RPKM)), followed by HLA-C (lung, 371.9 and breast, 177.1 RPKM), then HLA-A (lung, 289.4 and breast, 162.0 RPKM; Fig. 2a). A wide range of HLA expression was observed across the three class I genes (Fig. 2a). In individuals with both lung and breast tissue samples, HLA gene expression was higher in lung tissue in 175/198 (88%), 186/204 (91%) and 190/203 (94%) of cases for HLA-A, HLA-B and HLA-C, respectively (Fig. 2b).

Significant HLA allelic imbalance (AIB) was pervasive: 273/388 (70%), 183/418 (44%) and 305/407 (75%) of normal lung tissue samples exhibited statistically significant AIB in expression in HLA-A, HLA-B and HLA-C, respectively, while 266/329 (81%), 178/343 (52%) and 255/328 (78%) of normal breast tissue samples exhibited statistically significant AIB expression in HLA-A, HLA-B and HLA-C, respectively (Fig. 2c and Supplementary Note).

This observed expression imbalance may be due in part to the combination of HLA alleles that an individual harbors. Consistently, we observed evidence of a relationship between allelic expression and the allele type, consistent with previous reports^{21,22} ($P < 2 \times 10^{-16}$, one-way analysis of variance; Fig. 2d). From the alleles for which we had >30 GTEx individuals with lung and/or breast samples, we found that in lung, HLA-A*03:01, HLA-B*15:01 and HLA-C*08:02 had the lowest expression across the three genes, while HLA-A*24:02, HLA-B*40:01 and HLA-C*04:01 had the highest. In breast, HLA-A*03:01, HLA-B*15:01 and HLA-C*07:02 had the lowest expression across the three genes, whereas HLA-A*24:02, HLA-B*08:01 and HLA-C*04:01 had the highest.

HLA alternative splicing in normal tissue samples

Given the role that HLA alternative splicing could have in HLA presentation, we first used MHC Hammer to investigate the prevalence of HLA alternative splicing in the GTEx normal tissue cohort. In our cohort, MHC Hammer identified complete exon skipping, partial exon skipping and partial intron retention in the HLA alleles, but we did not observe evidence for complete intron retention in any HLA allele (Extended Data Fig. 4 and Supplementary Note).

Alternative splicing in the HLA alleles was frequent in the GTEx normal tissue cohort; 466/483 (97%) of normal lung and 339/392 (87%) of normal breast samples harbored at least one alternative splicing event. Exon 5 skipping was the most frequent event in both lung and breast tissue, occurring in 368/483 (76%) of lung and 270/392 (69%) of breast samples, followed by partial retention of intron 5, occurring in 271/483 (56%) of lung and 180/392 (46%) of breast tissue samples (Fig. 2e,f). Skipping of exon 5 has been shown to result in a soluble HLA molecule due to the absence of the transmembrane domain^{10–12}. We also observed HLA alternative splicing events in exons or introns 2, 3 or 4, which could result in an unstable HLA molecule potentially unable to present antigens to the immune system^{8,9,13} in 267/483 (55%) of normal lung tissue samples and 147/392 (38%) of normal breast tissue samples (Fig. 2e,f).

To estimate the relative abundance of the novel (alternatively spliced) transcripts, we estimated a 'novel transcript proportion' (Supplementary Note). Most splicing events occurred with a low novel transcript proportion, with 1,667/1,863 (90%) events in the lung and 956/1,118 (86%) events in the breast occurring with a novel transcript proportion of less than 0.1 (Fig. 2g).

We next compared the splicing events in breast and lung tissues from the same patient. In total, 43% of all alternative splicing events occurred in both breast and lung tissue, while 46% occurred only in the lung sample and just 11% occurred only in the breast sample. When restricting to events that occurred in both tissues from the same individual, we observed a high concordance between the novel transcript proportions (Fig. 2h). However, for the majority of the alternative splicing events that introduced a premature termination codon (PTC), the novel transcript proportion was higher in lung than in breast tissue (Extended Data Fig. 5). This observation could possibly be driven by

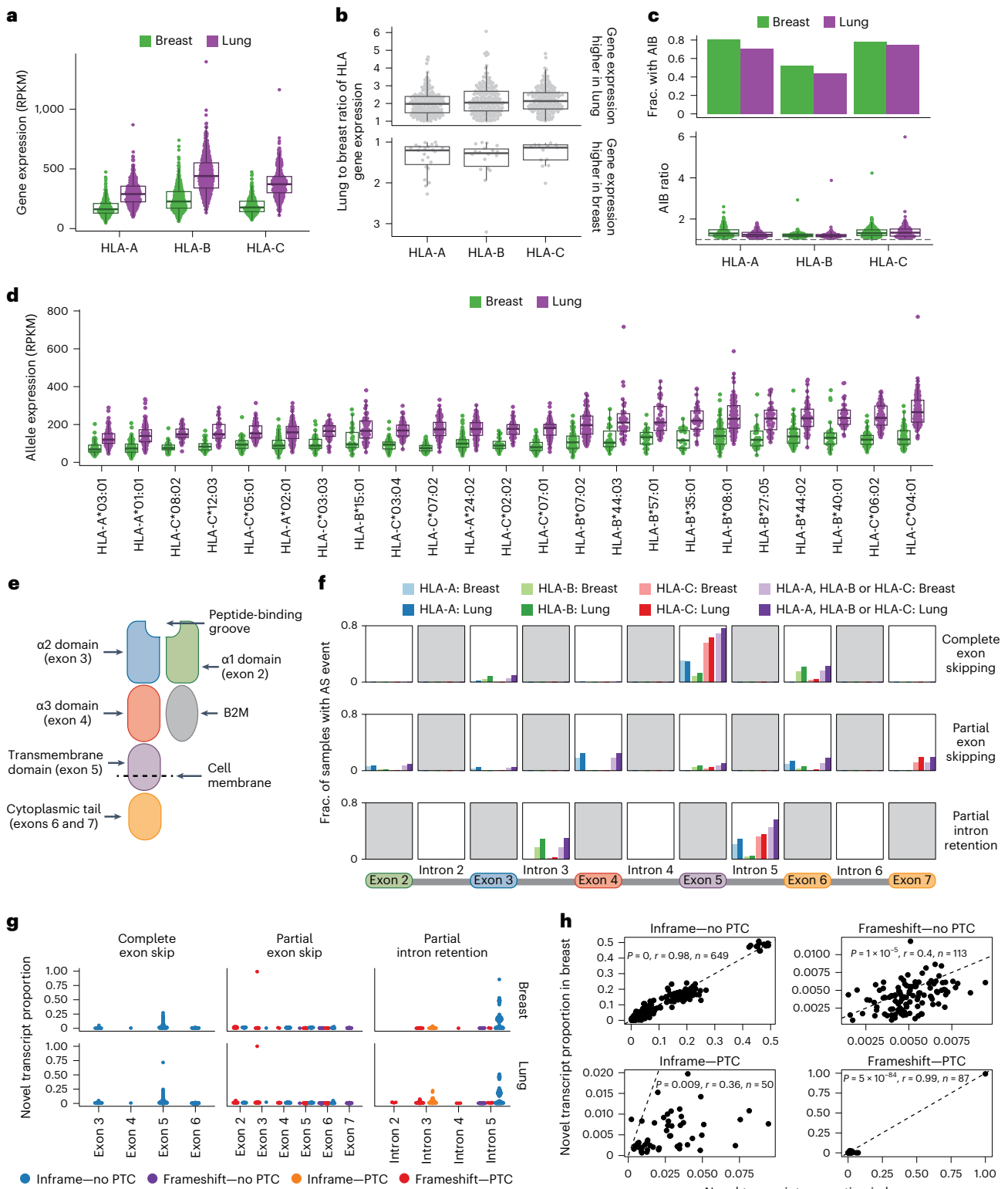


Fig. 2 | HLA expression is variable in normal tissue. **a**, Gene level expression in the GTEx lung and breast normal tissue samples for HLA-A, HLA-B and HLA-C (lung, $n = 483$ and breast, $n = 392$). **b**, The ratio of the lung to breast HLA gene expression ($n = 238$ patients with both a lung and breast sample). **c**, The fraction of tumor-adjacent normal samples with RNA AIB (top) and for the samples with AIB, the AIB ratio (bottom; lung, $n = 440$ and breast, $n = 380$). **d**, The allelic expression per allele type in lung and breast tissue. Only alleles with at least 30 lung and/or breast samples are included in this analysis (lung, $n = 465$ and breast, $n = 377$). **e**, The structure of the class I HLA molecule. **f**, Rates of alternative splicing in the GTEx normal lung and breast samples for the exons and introns

shown along the bottom. **g**, The novel transcript proportion of the alternative splicing events. **h**, The relationship between the novel transcript proportion in the lung and breast tissues of the same GTEx individual. The P value and correlation coefficient (r) in **h** are calculated using Pearson's method. Boxplots in **a–d** show the median and first and third quartiles, and whiskers extend to $1.5 \times$ IQR above and below the IQR. IQR, interquartile range; AIB, allelic imbalance; AS, alternative splicing; PTC, premature termination codon; RPKM, reads per kilobase per million.

differences in the rate of nonsense mediated decay (NMD) in different tissues^{23,24}.

Finally, we investigated whether certain alleles were enriched for specific alternative splicing events. We focused on alternative splicing events in the 23 HLA alleles that were present in >30 GTEx individuals with breast and/or lung tissue. In lung, 10/105 alternative splicing events in these 23 alleles occurred in more than 50% of samples with the allele. In breast, 8/81 of alternative splicing events occurred in more than 50% of samples with the allele. For example, inframe complete exon 5 skipping in HLA-C*04:01:01:01 occurred in 100% of breast and lung samples (lung: 94/94, breast: 64/64) with this allele, which supports the results described in a previous study¹². We also observed inframe partial intron 5 retention in HLA-C*03:04:01:01 in 100% of breast and lung samples with this allele (lung, 70/70 and breast, 57/57; Supplementary Table 1).

These data suggest that the HLA alleles are subject to widespread expression imbalance and alternative splicing in normal tissue and that total HLA gene expression is strongly influenced by the combination of HLA alleles that a person harbors. These data emphasize the importance of controlling for HLA allelic expression in normal tissue when assessing transcriptional alterations in tumors.

HLA genomic disruption in lung and breast cancer

In the TRACERx421 cohort, LOH of the class I HLA genes was frequent, occurring in 75/235 (32%) of lung adenocarcinoma (LUAD), 76/132 (58%) of lung squamous cell carcinoma (LUSC) and 13/44 (30%) of other NSCLC histological subtype primary tumors, consistent with our previous findings⁷ (Extended Data Fig. 6). In keeping with this, the rate of HLA LOH in the TCGA lung cohort was 65/245 (27%) for LUAD tumors and 104/267 (39%) for LUSC tumors (Extended Data Fig. 6). In TCGA, triple-negative breast cancer (TNBC) had the highest rate of HLA LOH (17/61 (28%)), followed by estrogen receptor negative (ER-; 7/32 (22%)) and estrogen receptor positive (ER+; 60/402 (15%); Extended Data Fig. 6). By contrast, high-impact damaging mutations in the HLA genes were relatively rare, occurring in only 5/411 (1.2%) tumors in the TRACERx421 cohort, 2/514 (0.4%) tumors in the TCGA breast cohort and were not observed in the TCGA lung cohort.

Transcriptional repression of class I HLA alleles in tumors

We next investigated whether there was evidence of additional disruption of HLA alleles through transcriptional repression in tumor regions. Given the heterogeneity observed in normal HLA allelic expression, we measured tumor HLA repression with reference to the patient-matched tumor-adjacent normal sample (Supplementary Note). We were able to evaluate transcriptional repression in 49 LUAD and 29 LUSC tumors in the TRACERx421 cohort and in 13 LUAD, 27 LUSC and 34 ER+ breast cancers from the TCGA cohort. We did not detect any high-impact damaging HLA mutations in any tumor with a tumor-adjacent normal sample.

We identified extensive transcriptional repression of the HLA alleles that could not be explained by LOH or damaging mutations in both the lung and breast tumors (Fig. 3a,b and Supplementary Fig. 2). In the TRACERx421 cohort, 30/49 (61%) of LUAD and 22/29 (76%) of LUSC tumors harbored transcriptional repression of at least one HLA allele not caused by LOH. Taken together, just 13/49 (27%) of LUAD and 2/29 (7%) of LUSC tumors exhibited no LOH or repression in any class I HLA gene (Fig. 3b). These results were consistent in the TCGA LUAD and LUSC cohorts (Supplementary Fig. 2). In contrast, 19/34 (56%) of ER+ breast cancers exhibited no damaging mutations, LOH or repression in any class I HLA gene (Fig. 3b).

HLA genomic biallelic loss (that is, homozygous deletion) was an uncommon event, occurring in only 11/411 (3%) of TRACERx421 NSCLCs, 7/512 (1%) of TCGA NSCLCs and 17/514 (3%) of TCGA breast cancers. To investigate biallelic transcriptional repression, we restricted our analysis to HLA genes with no evidence for genomic alterations. Biallelic transcriptional repression of a given HLA gene was relatively frequent,

occurring in 24/43 (56%), 11/16 (69%) and 9/32 (28%) of LUAD, LUSC and ER+ breast tumors (Fig. 3c). However, while homozygous deletion will necessarily impact both alleles equally, we found evidence of unequal biallelic repression of alleles, with 14/24 (58%), 8/11 (73%) and 4/9 (44%) of LUAD, LUSC and ER+ tumors harboring at least one HLA gene with AIB in the tumor but not the normal (or vice versa).

The impact of HLA disruption on neoantigen presentation

To investigate the impact of HLA LOH and transcriptional repression on the predicted number of neoantigens presented to the immune system, we quantified, for each tumor region, the number of different alleles when considering (1) neither LOH nor repression, (2) LOH or (3) LOH and repression. When accounting for LOH and repression, 39/132 (30%) of LUAD tumor regions, 3/90 (3%) of LUSC tumor regions and 18/36 (50%) of ER+ breast tumors had all six intact HLA alleles, while 9/132 (7%) of LUAD tumor regions, 18/90 (20%) of LUSC tumor regions and 3/36 (8%) of ER+ breast tumors had all six alleles disrupted at the genomic and transcriptomic levels (Fig. 3d). On average, 28.2% and 52.3% of putative neoantigens were predicted to bind exclusively to alleles subject to LOH or repression in LUAD and LUSC, respectively (Extended Data Fig. 7).

Mechanisms of HLA repression

The predominant modulators of HLA class I transcription are the NOD-like receptor (NLR) proteins NLRC5 and CIITA. The HLA promoter region also contains the tumor necrosis factor (TNF)-stimulated promoter site, EnhA and the IFNG-stimulated response element, ISRE²⁵. In tumors without any genomic HLA disruption, we observed a significant positive correlation (Pearson's $r \geq 0.3$ and $P \leq 0.01$), between total HLA expression and the expression of *NLRC5* and *CIITA* in LUAD, LUSC and ER+ tumors from both the TRACERx421 and TCGA cohorts, as well as between total HLA expression and *IFNG* expression in the TRACERx LUAD and TCGA LUAD and LUSC cohorts. We only observed a significant positive correlation between total HLA expression and *TNF* in the TCGA LUSC cohort (Extended Data Fig. 8).

We next investigated whether there was a relationship between tumor-normal changes in the expression of these genes and the likelihood of the tumor region having allelic repression unexplained by genomic disruption. In LUAD, LUSC and ER+ breast cancer, samples with allelic transcriptional repression had a significantly lower tumor-to-normal ratio of *NLRC5* than those without transcriptional repression. The same was true for *CIITA* and *IFNG* in LUAD and LUSC tumors, but not ER+ breast tumors (Fig. 3e). We only observed a significant relationship with *TNF* in LUADs (Supplementary Fig. 3).

Previous work has identified methylation as a mechanism that can influence HLA allelic expression^{21,26}. To investigate the role of hypermethylation in HLA transcriptional repression, we used methylation array data from the TCGA cohorts. We observed distinct patterns of methylation across the HLA genes in the LUAD, LUSC and breast tumors, with the gene body having the highest methylation in HLA-A and HLA-B in both tumor and normal tissues. In HLA-C, the region 1,500–200 bp upstream of the transcriptional start site (termed TSS1500) had the highest level of methylation, followed by the gene body, in both the tumor and normal tissues (Supplementary Fig. 4). We observed a significant negative correlation (Pearson's $r \leq -0.3$ and $P \leq 0.01$) between HLA-B gene expression and the degree of methylation in both the TSS1500 and the gene body region in the TCGA LUAD, LUSC and ER+ tumor samples (Supplementary Fig. 5–7).

Taken together, these data suggest that in the LUAD, LUSC and breast tumors, changes in the expression of *NLRC5*, *CIITA*, *IFNG* as well as hypermethylation may play a role in the repression of HLA.

HLA alternative splicing in breast and lung cancer

Given the pervasive nature of alternative splicing in normal tissue, we evaluated whether splicing events occurred at significantly higher

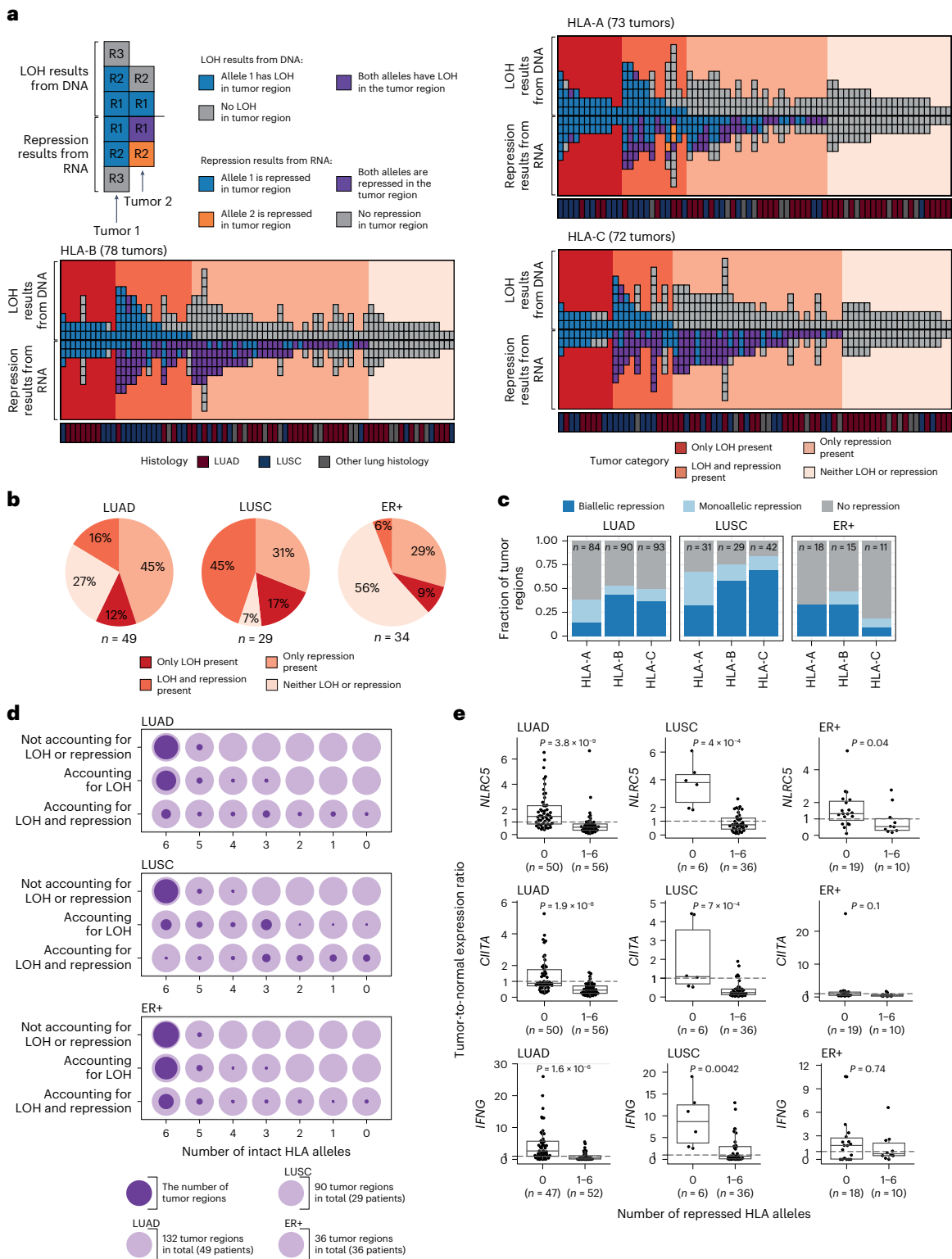


Fig. 3 | Transcriptional repression of the HLA genes in lung and ER+ breast cancer. a, Each column represents a tumor, and each box represents a region from that tumor. Each region appears as two boxes. The first top box (above line) is coloured by the regional HLA LOH status, and the second bottom box (mirrored below line), coloured by whether the region has transcriptional repression of the same allele that is lost in the DNA (blue), the alternate allele (allele 2, orange) or both alleles (purple). Only tumors with a patient-matched tumor-adjacent normal sample are included in this figure. None of these tumors had a high-impact damaging HLA mutation. **b**, The fraction of tumors with either only HLA LOH, only repression (unexplained by genomic alterations), both HLA LOH and repression (unexplained by genomic alterations) or no HLA LOH or

repression. **c**, The frequency of monoallelic and biallelic repression events in tumor regions without genomic HLA alterations. **d**, The total number of intact alleles when accounting for alleles disrupted by LOH and repression. The lighter circle indicates the number of tumor regions in total, and the superimposed darker circle indicates the number of tumor regions in the given category. **e**, The relationship between the tumor-to-normal ratio of *NLRCS*, *CIITA* and *IFNG* expression and the number of transcriptionally repressed alleles in the tumor region. The *P* value in **e** is derived from a two-sided Wilcoxon test. Boxplots in **e** show median and first and third quartiles, and whiskers extend up to 1.5× IQR above and below the IQR. LOH, loss of heterozygosity.

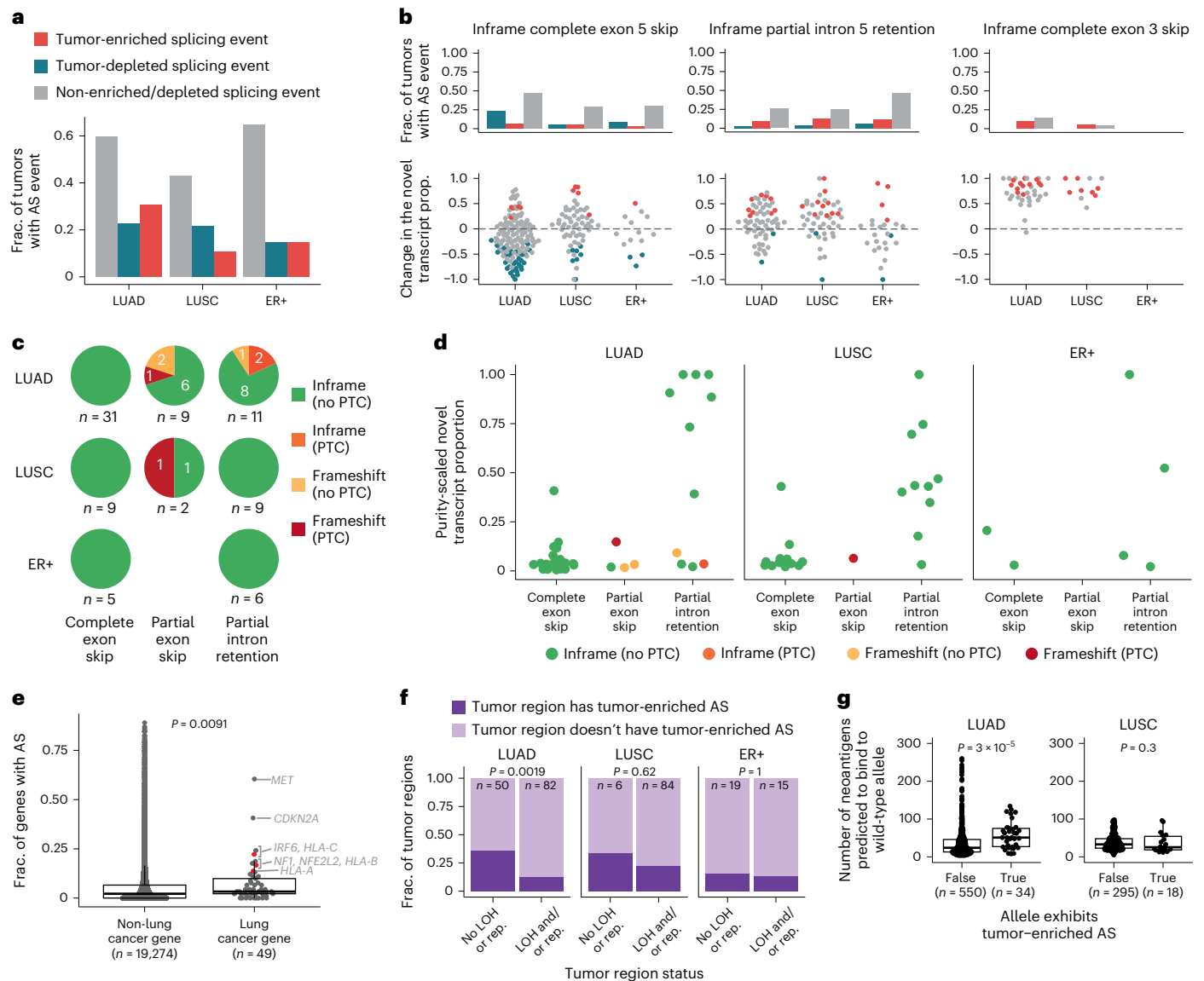


Fig. 4 | HLA alternative splicing in lung and breast tumors. **a**, The fraction of tumors that exhibit tumor-enriched or tumor-depleted HLA alternative splicing events in the TRACERx (LUAD and LUSC) and TCGA (ER+) cohorts. **b**, The three most frequent HLA alternative splicing events are shown. The fraction of tumors that exhibit the event is shown at the top and the tumor-to-normal change in the novel transcript proportion is shown at the bottom. The legend is shown in **a**. **c**, The predicted consequences of the tumor-enriched and tumor-depleted alternative splicing events. **d**, The purity-scaled novel transcript proportion for the tumor-enriched alternative splicing events. **e**, The fraction of TRACERx tumors that exhibit alternative splicing events across all protein-coding genes, split by whether the gene is classified as a lung cancer gene or not. The three HLA

genes are shown in red. **f**, The fraction of tumor regions that do/do not have LOH or repression and do/do not have tumor-enriched alternative splicing events. **g**, The neoantigen count for each allele is split by whether the allele exhibits tumor-enriched alternative splicing or not. Only alleles with no genomic disruption were included. *P* values in **e** and **g** are derived from a two-sided Wilcoxon test. *P* value in **f** is derived from Fisher's exact test. Boxplots in **e** and **g** show median and first and third quartiles, and whiskers extend up to 1.5× IQR above and below the IQR. AS, alternative splicing; LOH, loss of heterozygosity; TSG, tumor suppressor gene; PTC, premature termination codon; rep, repression.

frequency in the tumor ('tumor-enriched') or, conversely, at significantly lower frequency in the tumor ('tumor depleted'; Supplementary Note). Both tumor-enriched and tumor-depleted HLA alternative splicing events were common, with 30.6%, 10.7% and 14.7% of LUAD, LUSC and ER+ breast tumors harboring at least one tumor-enriched alternative splicing event and 22.6%, 21.4% and 14.7% of LUAD, LUSC and ER+ tumors harboring at least one tumor-depleted alternative splicing event (Fig. 4a).

The most frequent tumor-enriched alternative splicing events included inframe partial intron 5 retention and inframe complete exon 3 skipping (Fig. 4b). Changes to the sequence that encodes exon 3

could result in altered peptide binding or an unstable HLA molecule^{8,13}. In contrast, inframe complete exon 5 skipping was observed more frequently as a tumor-depleted event in LUAD and ER+ tumors (Fig. 4b). Skipping of exon 5 has been demonstrated in previous studies to result in a soluble HLA molecule^{10–12}. For each alternative splicing event, the tumor-to-normal change in the novel transcript proportion is shown (Supplementary Note and Fig. 4b).

The introduction or deletion of nucleotide sequences due to alternative splicing could result in a frameshift and/or the introduction of a PTC in the resulting transcript. In LUAD, LUSC and ER+ tumors, all complete exon skipping events were inframe (Fig. 4c). In contrast,

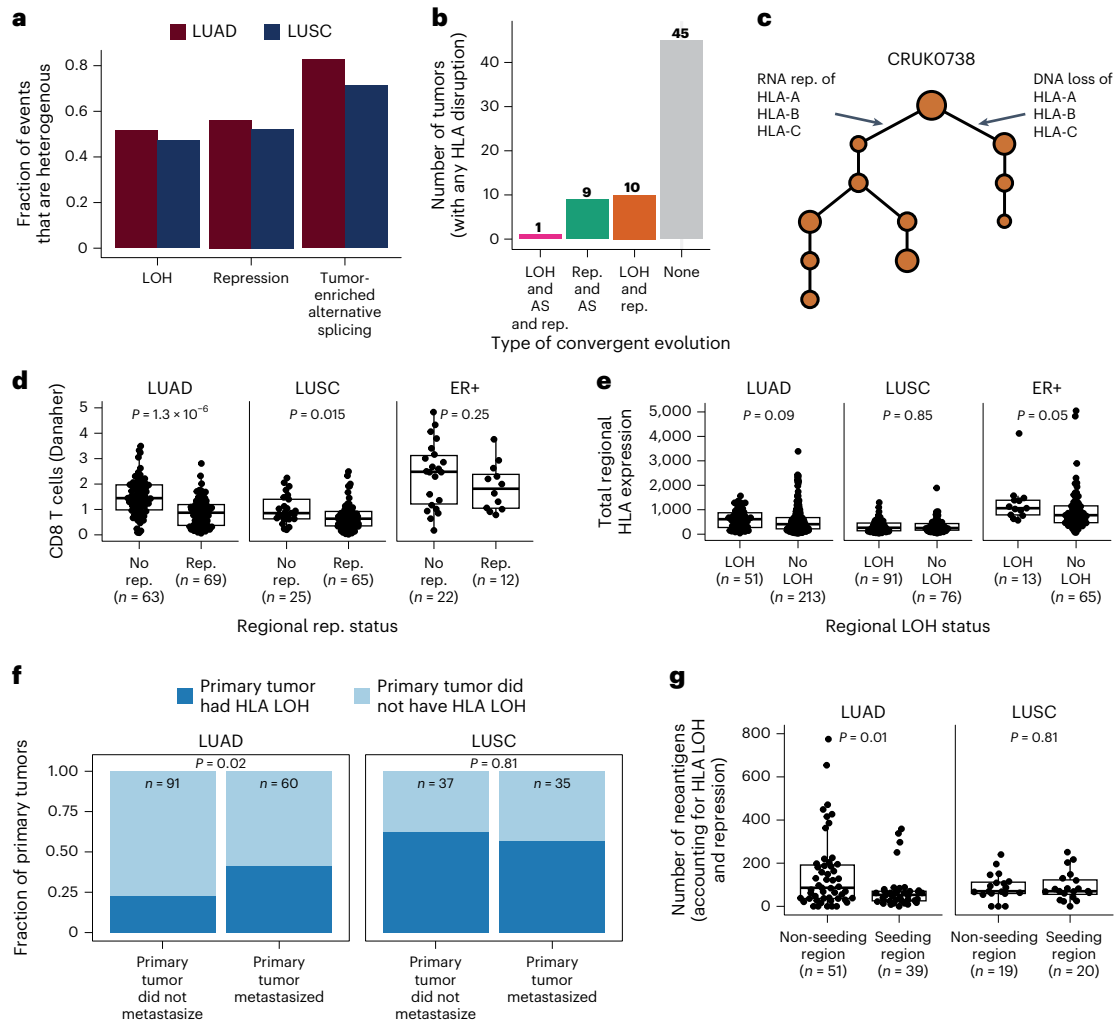


Fig. 5 | The role of HLA disruption in tumor evolution. a, The heterogeneity of HLA LOH, repression and tumor-enriched alternative splicing events. **b,c**, Overview (**b**) and example (**c**) of convergent evolution, where the same HLA allele is disrupted via different mechanisms in different regions of the same tumor. **d**, The relationship between the presence of repression and the amount of CD8 T cell infiltration. **e**, Tumor regions with and without HLA LOH have similar levels of total HLA expression. **f**, LUAD tumors that have HLA LOH are more likely

to metastasize. **g**, When accounting for LOH and repression, LUAD regions that seeded a metastasis have a lower neoantigen count than those that did not. *P* values in **d**, **e** and **g** are derived from a two-sided Wilcoxon test. *P* value in **f** is derived from Fisher's exact test. Boxplots in **d**, **e** and **g** show median and first and third quartiles, and whiskers extend up to $1.5 \times$ IQR above and below the IQR. LOH, loss of heterozygosity; rep, repression; AS, alternative splicing.

we observed partial exon skipping events and partial intron retention events that were inframe, or that resulted in a frameshift, with and without the introduction of a PTC (Fig. 4c).

Given that tumor samples reflect an admixture of cancer cells and non-cancer cells, to estimate the fraction of alternatively spliced transcripts in the cancer cells, we scaled the novel transcript proportion of the tumor-enriched events by the estimated purity of the tumor region (Supplementary Note). Although there were outliers, the purity-scaled novel transcript proportion was less than 0.25 in most cases (mean = 0.24 and range = 0.006–1; Fig. 4d). These data suggest either one or both of the following are occurring: within each cancer cell, both the canonical and novel transcripts are being transcribed, or only a subset of cancer cells harbor the novel transcript.

To further evaluate the rate of tumor-enriched alternative splicing observed in HLA alleles and whether this is higher or lower than expected, we considered the rate of somatic alternative splicing across all protein-coding genes (Methods). We found that lung cancer genes had a higher rate of alternative splicing than other protein-coding genes ($P = 9.1 \times 10^{-3}$; Fig. 4e). In addition, from the set of 49 lung cancer genes, we found that HLA-C had the fourth, HLA-B had

the seventh and HLA-A had the eighth highest frequency of alternative splicing (Fig. 4e).

Consistent with the selection of alternative splicing events, we observed that LUAD tumor regions without HLA LOH or repression were enriched for tumor-enriched alternative splicing events (LUAD, $P = 1.9 \times 10^{-3}$) compared to regions that harbored either HLA LOH or repression (Fig. 4f). We did not see this enrichment with LUSC or ER+ tumors. This suggests that in LUAD, tumor-enriched alternative splicing may offer an alternative means to disrupt HLA presentation during tumor evolution.

To further investigate the importance of HLA alternative splicing in tumor evolution, we compared the total number of neoantigens predicted to bind to alleles with or without tumor-enriched HLA alternative splicing. We first quantified the number of neoantigens predicted to bind to the intact HLA alleles for each tumor region. We then compared the neoantigen count in alleles that exhibited tumor-enriched alternative splicing versus those that did not, excluding alleles with HLA LOH. HLA alleles exhibiting tumor-enriched HLA alternative splicing in LUAD tumors were associated with a higher neoantigen count compared to alleles without evidence of tumor-enriched splicing ($P = 3 \times 10^{-5}$;

Fig. 4g), suggesting that in LUAD tumors, tumor-enriched alternative splicing of HLA alleles may be selected to reduce antigen presentation. We did not observe this enrichment in LUSC tumors.

HLA disruption and tumor evolution

To understand when HLA LOH, transcriptional repression and somatic alternative splicing occur during NSCLC evolution, we considered the heterogeneity of these events using the multiregion TRACERx cohort. We defined an HLA disruption event as ubiquitous if it occurred in all of the primary tumor regions, and heterogeneous otherwise. In both LUAD and LUSC tumors, tumor-enriched alternative splicing events were the most heterogeneous (LUAD = 82.6% and LUSC = 71.4%), followed by repression (LUAD = 56.2% and LUSC = 52.2%) and then LOH events (LUAD = 51.6% and LUSC = 47.1%; Fig. 5a).

In 20/65 (30.8%) of TRACERx LUAD and LUSC tumors with HLA disruption, we observed convergence upon disruption of the same allele through alternative mechanisms, with genomic loss, transcriptional repression and/or alternative splicing of the same allele occurring in different regions of the same tumor. We observed ten tumors with convergence upon genomic loss and transcriptional repression of the same allele in separate regions, nine tumors with transcriptional repression and alternative splicing of the same allele in separate regions and one tumor with genomic loss, transcriptional repression and alternative splicing of the same allele in different regions (Fig. 5b,c). Conceivably, this could either reflect positive selection within individual tumors or be a consequence of the high rate of HLA disruption through diverse mechanisms.

The tumor microenvironment can shape tumor evolution¹. We therefore investigated the relationship between the immune infiltrate and the presence of HLA disruption, using the Danaher *in silico* immune deconvolution method to estimate the amount of CD8 T cell infiltrate²⁷. We observed a significant relationship between total HLA expression and CD8 T cell infiltrate (LUAD, $P = 1.6 \times 10^{-27}$ and $r = 0.48$; LUSC, $P = 3.7 \times 10^{-15}$ and $r = 0.44$; ER+, $P = 1.3 \times 10^{-24}$ and $r = 0.42$; Supplementary Fig. 8). We observed that LUAD and LUSC tumor regions with allelic HLA transcriptional repression had lower levels of infiltrating CD8 T cells compared to those without (LUAD, $P = 1.3 \times 10^{-6}$; LUSC, $P = 0.015$ and ER+, $P = 0.25$; Fig. 5d). Conversely, HLA alternative splicing was associated with elevated CD8 T cell levels in LUAD tumors (LUAD, $P = 3 \times 10^{-6}$; LUSC, $P = 0.32$ and ER+, $P = 0.54$; Extended Data Fig. 9). No clear relationship between HLA LOH and total HLA expression in either LUAD, LUSC or ER+ tumors was observed, indicating dosage compensation may occur following allelic HLA copy number loss (Fig. 5e).

Finally, we endeavored to understand whether disruption of the HLA alleles through LOH, repression or alternative splicing might have a role in the evolution of lung cancer metastasis. We found that LUAD tumors harboring HLA LOH were more likely to metastasize than those without HLA LOH (LUAD, $P = 0.02$ and LUSC, $P = 0.81$; Fig. 5f). To investigate this further, we considered the neoantigen burden of primary tumor regions that seeded metastasis compared to those that did not, with the metastasis-seeding regions being defined in our previously published work¹⁵. We found that the standard neoantigen burden did not distinguish seeding from non-seeding regions (Extended Data Fig. 10a). However, when we restricted our neoantigen count to only include neoantigens predicted to bind to intact HLA alleles, not subject to loss or repression, we observed that LUAD tumor regions that seeded metastasis had a lower effective neoantigen burden than those that did not (LUAD, $P = 0.01$ and LUSC, $P = 0.81$; Fig. 5g), which was not the case when we only considered HLA loss (Extended Data Fig. 10b). Taken together, these data suggest that disruption of the HLA alleles could have an important role in tumor metastasis.

Discussion

Neoantigen presentation via HLA molecules is crucial to achieving an antitumor immune response. Previous studies have illustrated that

different mechanisms of HLA disruption are common across cancers^{4–7}. Here we developed MHC Hammer, a tool to investigate the prevalence of four mechanisms of genomic and transcriptomic disruption of the HLA alleles in lung and breast cancer—mutations, LOH, repression and alternative splicing.

While damaging HLA mutations were rare in our cohorts, LOH, repression and tumor-enriched alternative splicing of the HLA alleles were pervasive. From the patients with tumor-adjacent normal samples, just 27% of LUAD, 7% of LUSC and 56% of ER+ tumors had no HLA disruption, while 30.6%, 10.7% and 14.7% of LUAD, LUSC and ER+ tumors exhibited tumor-enriched alternative splicing events. The lower rate of HLA LOH and repression observed in ER+ breast tumors may reflect the lower tumor mutational burden (TMB) in ER+ breast tumors compared to NSCLC²⁸.

We observed differences in the patterns of HLA disruption in the NSCLC tumors—LUSC tumors were characterized by almost universal HLA disruption, while LUAD tumors exhibited less frequent HLA disruption. In LUADs, we observed an enrichment for alternative splicing in alleles without LOH or repression, a higher likelihood of tumor-enriched alternative splicing in alleles with a higher neoantigen burden and finally an association between HLA LOH and metastasis. This may reflect different selective pressures in these cancer types and the propensity for HLA disruption through diverse mechanisms.

One limitation of our method is that it requires a patient-matched tumor-adjacent normal tissue sample to determine HLA repression and tumor-enriched alternative splicing. This is due to the high variability observed in HLA allelic expression and the high prevalence of HLA alternative splicing in the normal tissue samples.

Alternative splicing of the class I HLA alleles has been observed in non-cancer tissue cohorts and in cancer cell lines^{8–13}. However, HLA alternative splicing in large cohorts of normal and tumor tissue has not been described before, due in part to the lack of a high-throughput bioinformatics tool capable of measuring HLA alternative splicing.

HLA alternative splicing affecting exons or introns 2–4 could result in an unstable HLA molecule. For example, partial exon 3 skipping in an HLA-A allele in non-cancer tissue has been shown to result in the absence of cell-surface expression¹³. An HLA-A allele with complete exon 3 skipping continued to be expressed on the cell surface but as an immature glycoprotein unable to present peptides⁸. This immature molecule could potentially act as a decoy allele by inhibiting NK cells via its receptor ligands without presenting neoantigens to CD8 T cells. Alternative splicing resulting in exon 5 skipping has been shown to result in a soluble HLA allele^{10–12}. Persistent presentation of neoantigens via soluble HLA molecules to the T cell receptor, without costimulatory or accessory signals, could lead to immune tolerance or T cell exhaustion. It has been shown that soluble class I HLA molecules can induce apoptosis in CD8 T cells and NK cells²⁹.

The majority of the detected tumor-enriched alternative splicing events were present with a purity-adjusted novel transcript proportion <0.25. This could reflect NMD; PTC-induced NMD has been shown to reduce mRNA levels by up to 90% in a study of an HLA-A allele²³. Alternatively, HLA alternative splicing may be a transient event, or the alternative splicing observed in the lung and breast tumors in this study may simply reflect transcriptional noise. Therefore, further work is required to establish the role of alternative splicing in lung and breast cancer, as well as investigate its prevalence in other cancer types.

It is possible that the underlying mechanisms of HLA repression and alternative splicing events are epigenetic. Supporting this, we found a strong link between methylation and expression of the HLA genes. This could have important clinical implications, as previous studies have illustrated the reversible nature of HLA epigenetic modifications²⁶ and the importance of this reversibility in immunotherapy response³⁰.

Further work is warranted to explore the extent to which HLA alternative splicing and repression represent a pan-cancer immune evasion

mechanism. As more pre-therapy and post-therapy data emerge, it will be possible to investigate the extent to which HLA alternative splicing and repression develop during treatment and at immune-therapy resistance and the extent to which these processes might inform therapeutic strategies.

Our results may also have implications for vaccine- and T cell-based therapeutic approaches, which seek to exploit neoantigens. Our results suggest that it may be important to consider not just whether putative neopeptides bind the repertoire of HLA alleles but also the copy number, expression and splicing characteristics of each allele. Indeed, MHC Hammer may be used to help determine which set of predicted neoantigens are most likely to elicit an effective T cell response.

In conclusion, MHC Hammer enables accurate estimation of allele-specific HLA disruption, revealing that it is a common feature of NSCLC and ER+ breast cancer that facilitates immune escape and cancer evolution.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-024-01883-8>.

References

- Rosenthal, R. et al. Neoantigen-directed immune escape in lung cancer evolution. *Nature* **567**, 479–485 (2019).
- Bentham, R. et al. Using DNA sequencing data to quantify T cell fraction and therapy response. *Nature* **597**, 555–560 (2021).
- Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48–61 (2015).
- Schaafsma, E., Fugle, C. M., Wang, X. & Cheng, C. Pan-cancer association of HLA gene expression with cancer prognosis and immunotherapy efficacy. *Br. J. Cancer* **125**, 422–432 (2021).
- Shukla, S. A. et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* **33**, 1152–1158 (2015).
- Momburg, F. et al. Loss of HLA-A,B,C and de novo expression of HLA-D in colorectal cancer. *Int. J. Cancer* **37**, 179–184 (1986).
- McGranahan, N. et al. Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell* **171**, 1259–1271 (2017).
- Dai, Z.-X., Zhang, G.-H., Zhang, X.-H., Zhu, J.-W. & Zheng, Y.-T. A splice variant of HLA-A with a deletion of exon 3 expressed as nonmature cell-surface glycoproteins forms a heterodimeric structure with full-length HLA-A. *Hum. Immunol.* **75**, 234–238 (2014).
- Wang, Z., Marincola, F. M., Rivoltini, L., Parmiani, G. & Ferrone, S. Selective histocompatibility leukocyte antigen (HLA)-A2 loss caused by aberrant pre-mRNA splicing in 624MEL28 melanoma cells. *J. Exp. Med.* **190**, 205–215 (1999).
- Krangel, M. S. Secretion of HLA-A and -B antigens via an alternative RNA splicing pathway. *J. Exp. Med.* **163**, 1173–1190 (1986).
- Yang, D. & Le, J. Targeted amplification of alternatively spliced transcripts of major histocompatibility complex class I heavy chain. *J. Immunol. Methods* **176**, 265–270 (1994).
- Ehlers, F. A. I. et al. Polymorphic differences within HLA-C alleles contribute to alternatively spliced transcripts lacking exon 5. *HLA* **100**, 232–243 (2022).
- Reinders, J. et al. Identification of HLA-A*0111N: a synonymous substitution, introducing an alternative splice site in exon 3, silenced the expression of an HLA-A allele. *Hum. Immunol.* **66**, 912–920 (2005).
- Martínez-Ruiz, C. et al. Genomic–transcriptomic evolution in lung cancer and metastasis. *Nature* **616**, 543–552 (2023).
- Al Bakir, M. et al. The evolution of non-small cell lung cancer metastases in TRACERx. *Nature* **616**, 534–542 (2023).
- Frankell, A. M. et al. The evolution of lung cancer and impact of subclonal selection in TRACERx. *Nature* **616**, 525–533 (2023).
- Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
- Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
- Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
- GTEX Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
- Ramsuran, V. et al. Epigenetic regulation of differential HLA-A allelic expression levels. *Hum. Mol. Genet.* **24**, 4268–4275 (2015).
- Johansson, T., Yohannes, D. A., Koskela, S., Partanen, J. & Saavalainen, P. HLA RNA sequencing with unique molecular identifiers reveals high allele-specific variability in mRNA expression. *Front. Immunol.* **12**, 629059 (2021).
- Watanabe, Y., Magor, K. E. & Parham, P. Exon 5 encoding the transmembrane region of HLA-A contains a transitional region for the induction of nonsense-mediated mRNA decay. *J. Immunol.* **167**, 6901–6911 (2001).
- Rivas, M. A. et al. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348**, 666–669 (2015).
- Carey, B. S., Poulton, K. V. & Poles, A. Factors affecting HLA expression: a review. *Int. J. Immunogenet.* **46**, 307–320 (2019).
- Rodems, T. S. et al. Reversible epigenetic alterations regulate class I HLA loss in prostate cancer. *Commun. Biol.* **5**, 897 (2022).
- Danaher, P. et al. Gene expression markers of tumor infiltrating leukocytes. *J. Immunother. Cancer* **5**, 18 (2017).
- Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
- Contini, P. et al. Soluble HLA-A,-B,-C and -G molecules induce apoptosis in T and NK CD8⁺ cells and inhibit cytotoxic T cell activity through CD8 ligation. *Eur. J. Immunol.* **33**, 125–134 (2003).
- Garrido, C. et al. Immunotherapy eradicates metastases with reversible defects in MHC class I expression. *Cancer Immunol. Immunother.* **60**, 1257–1268 (2011).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024

¹Cancer Evolution and Genome Instability Laboratory, The Francis Crick Institute, London, UK. ²Cancer Genome Evolution Research Group, Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK. ³Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK. ⁴Cancer Immunology Unit, Immune Regulation and Tumour Immunotherapy Laboratory, Research Department of Haematology, University College London Cancer Institute, London, UK. ⁵Cancer Cell Biology Laboratory, Centro de Biología Celular y Biomedicina (CEBICEM), Facultad de Medicina y Ciencia, Universidad San Sebastián, Santiago, Chile. ⁶Centro Ciencia & Vida, Fundación Ciencia & Vida, Santiago, Chile. ⁷Department of Molecular Oncology and Immunology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. ⁸Oncode Institute, Utrecht, The Netherlands. ⁹Tumour Immunogenomics and Immunosurveillance Laboratory, University College London Cancer Institute, London, UK. ¹⁰Division of Research, Peter MacCallum Cancer Centre, Melbourne, Australia. ¹¹Department of Pathology, ZAS Hospitals, Antwerp, Belgium. ¹²Department of Cellular Pathology, University College London Hospitals, London, UK. ¹³Cancer Genomics Laboratory, The Francis Crick Institute, London, UK. ¹⁴Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. ¹⁵Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. ¹⁶Cancer Metastasis Laboratory, University College London Cancer Institute, London, UK. ¹⁷Department of Medical Oncology, University College London Hospitals, London, UK. ¹¹⁴Immune Regulation and Immune Interactions Group, Centre for Immuno-Oncology, Nuffield Department of Medicine, University of Oxford, Oxford, UK. ✉e-mail: charles.swanton@crick.ac.uk; nicholas.mcgranahan.10@ucl.ac.uk

TRACERx Consortium

Heather Cheyne¹⁸, Mohammed Khalil¹⁸, Shirley Richardson¹⁸, Tracey Cruickshank¹⁸, Eric Lim^{19,20}, Hugo J. W. L. Aerts^{21,22,23}, Tom L. Kaufmann^{24,25}, Matthew R. Huska²⁶, Babu Naidu²⁷, Gareth A. Wilson¹, Rachel Rosenthal¹, Andrew Rowan¹, Chris Bailey¹, Claudia Lee¹, Emma Colliver¹, Katey S. S. Enfield¹, Mark S. Hill¹, Mihaela Angelova¹, Oriol Pich^{1,3}, Dhruva Biswas^{1,3,28}, Clare Puttick^{1,2,3}, Roberto Vendramin^{1,3,9}, Cian Murphy¹, Maria Zagorulya¹, Thomas P. Jones^{2,3}, Michelle M. Leung^{1,2,3}, Nicholas McGranahan^{2,3}, Carla Castignani^{13,29}, Elizabeth Larose Cadieux^{13,29}, Jeanette Kittel^{3,16}, Kerstin Haase^{3,16}, Kexin Koh^{3,16}, Rachel Scott^{3,16}, Gurdeep Matharu³⁰, Jacqui A. Shaw³⁰, Allan Hackshaw³¹, Camilla Pilotti³¹, Rachel Leslie³¹, Anne-Marie Hacker³¹, Sean Smith³¹, Aoife Walker³¹, Christopher Abbosh³, Corentin Richard³, Cristina Naceur-Lombardelli³, Francisco Gimeno-Valiente³, Krupa Thakkar³, Mariana Werner Sunderland³, Monica Sivakumar³, Nnennaya Kanu³, Ieva Usaite³, Sadegh Saghafinia³, Selvaraju Veeriah³, Sharon Vanloo³, Bushra Mussa³, Michalina Magala³, Elizabeth Keene³, Emilia L. Lim^{1,3}, James R. sM Black^{1,2,3}, Maise Al Bakir^{1,3}, Ariana Huebner^{1,2,3}, Kristiana Grigoriadis^{1,2,3}, Takahiro Karasaki^{1,3,16,32}, Alexander M. Frankell^{1,3}, Crispin T. Hiley^{1,3}, Sophia Ward^{1,3,33}, Sian Harries^{1,3,33}, Olivia Lucas^{1,3,34,35}, David A. Moore^{1,3,12}, Nicolai J. Birkbak^{1,3,36,37,38}, Carlos Martínez-Ruiz^{1,2,3}, Kerstin Thol^{2,3}, Robert Bentham^{2,3}, Wing Kin Liu^{3,16}, Abigail Bunkum^{3,16,34}, Sonya Hessey^{3,16,34}, Martin D. Forster^{3,17}, Siow Ming Lee^{3,17}, Mariam Jamal-Hanjani^{3,16,17}, Despoina Karagianni^{3,39}, Sergio A. Quezada^{3,4}, Supreet Kaur Bola^{3,4}, Kevin Litchfield^{3,9}, Charles Swanton^{1,2,17}, John Le Quesne^{40,41,42}, Khalid Abduljabbar⁴³, Catarina Veiga⁴⁴, Simone Zaccaria^{3,34}, Jonathan Tugwood^{45,46}, Caroline Dive^{45,46}, Zoltan Szallasi^{47,48,49}, Miklos Diossy^{47,48,50}, Teresa Marafioti¹², Elaine Borg¹², Mary Falzon¹², Reena Khuroy¹², Peter Van Looy^{13,14,15}, Karl S. Peggs^{51,52}, Gillian Price^{53,54}, Gary Royle⁵⁵, Charles-Antoine Collins-Fekete⁵⁵, Dionysis Papadatos-Pastos¹⁷, James Wilson¹⁷, Tanya Ahmad¹⁷, Sarah Benafif^{17,56}, Judith Cave⁵⁷, Keith M. Kerr^{54,58}, Thomas B. K. Watkins⁵⁹, Roberto Salgado^{10,11}, Alexander James Procter⁶⁰, Asia Ahmed⁶⁰, Magali N. Taylor⁶⁰, Arjun Nair^{60,61}, David Lawrence⁶², Davide Patrini⁶², Colin R. Lindsay^{46,63}, Fiona H. Blackhall^{46,63}, Yvonne Summers^{46,63}, Matthew G. Krebs⁶³, Emma Nye⁶⁴, Richard Kevin Stone⁶⁴, Hanyun Zhang⁶⁵, Jerome Nicod⁶⁶, Alan Kirk⁶⁷, Mo Asif⁶⁷, Rocco Bilancia⁶⁷, Nikos Kostoulas⁶⁷, Jennifer Whiteley⁶⁷, Mathew Thomas⁶⁷, Akshay J. Patel⁶⁸, David Chuter⁶⁹, Mairead MacKenzie⁶⁹, Roland F. Schwarz^{25,70}, Andrew Kidd⁷¹, Francesco Fraioli⁷², Paul Ashford⁷³, Zoltan Kaplar^{74,75}, Jonas Demeulemeester^{76,77}, Claire Wilson⁷⁸, Michael J. Shackcloth⁷⁹, Sam M. Janes⁸⁰, Neal Navani^{80,81}, Ricky M. Thakkar^{80,81}, Angela Leek⁸², Jack Davies Hodgkinson⁸², Nicola Totton⁸², Antonio Paiva-Correia⁸³, Stephan Beck²⁹, Miljana Tanic^{29,84}, Craig Dick⁸⁵, Lily Robinson⁸⁶, Peter Russell⁸⁶, Paulo De Sousa²⁰, Simon Jordan²⁰, Alexandra Rice²⁰, Hilgardt Raubenheimer²⁰, Harshil Bhayani²⁰, Lyn Ambrose²⁰, Anand Devaraj²⁰, Hemangi Chavan²⁰, Sofina Begum²⁰, Silviu I. Buderiu²⁰, Daniel Kaniu²⁰, Mpho Malima²⁰, Sarah Booth²⁰, Nadia Fernandes²⁰, Pratibha Shah²⁰, Chiara Proli²⁰, Andrew G. Nicholson^{20,87}, Ekaterini Boleti⁸⁸, Madeleine Hewish^{89,90}, Kevin G. Blyth^{91,92,93}, Jason F. Lester⁹⁴, Anshuman Chaturvedi^{46,95}, Pedro Oliveira^{46,95}, Katherine D. Brown^{46,95}, Mathew Carter^{46,95}, Alastair Magness⁹⁶, Clare E. Weedon⁹⁶, Eva Grönroos⁹⁶, Jacki Goldman⁹⁶, Mickael Escudero⁹⁶, Philip Hobson⁹⁶, Stefan Boeing⁹⁶, Tamara Denner⁹⁶, Vittorio Barbè⁹⁶, Wei-Ting Lu⁹⁶, William Hill⁹⁶, Yutaka Naito⁹⁶, Zoe Ramsden⁹⁶, George Kassiotis^{96,97}, Imran Noorani^{96,98}, Anca Grapa⁴³, Aiman Alzetani⁹⁹, Yinyin Yuan¹⁰⁰, Xiaoxi Pan¹⁰⁰, Jack French⁵⁶, Kayleigh Gilbert⁵⁶, Angela Dwornik¹⁰¹, Angeliki Karamani¹⁰¹, Benny Chain¹⁰¹, David R. Pearce¹⁰¹, Felip Gálvez-Cancino¹⁰¹, Georgia Stavrou¹⁰¹, Gerasimos-Theodoros Mastrokalos¹⁰¹, Helen L. Lowe¹⁰¹, Ignacio Garcia Matos¹⁰¹, James L. Reading¹⁰¹, John A. Hartley¹⁰¹, Kayalvizhi Selvaraju¹⁰¹, Kezhong Chen¹⁰¹, Leah Ensell¹⁰¹, Mansi Shah¹⁰¹, Maria Litovchenko¹⁰¹, Piotr Pawlik¹⁰¹, Samuel Gamble¹⁰¹, Seng Kuong Anakin Ung¹⁰¹, Victoria Spanswick¹⁰¹, Yin Wu¹⁰¹, Jayant K. Rane^{1,101}, Othman Al-Sawaf^{101,102}, Olga Chervova^{101,103}, Emilie Martinoni Hoogenboom³⁵, Fleur Monk³⁵, James W. Holding³⁵, Junaid Choudhary³⁵, Kunal Bhaktri³⁵, Pat Gorman³⁵, Robert C. M. Stephens³⁵, Maria Chiara Piscicella³⁵, Steve Bandula³⁵, Yien Ning Sophia Wong^{35,104}, Aya Osman¹⁰⁵, Mandeesh Sangha¹⁰⁵, Gerald Langman¹⁰⁵, Helen Shackleford¹⁰⁵, Madava Djearaman¹⁰⁵, Gary Middleton^{105,106}, Serena Chee¹⁰⁷, Patricia Georg¹⁰⁷, Amrita Bajaj¹⁰⁸, Apostolos Nakas¹⁰⁸, Azmina Sodha-Ramdeen¹⁰⁸, Mohamad Tufail¹⁰⁸, Molly Scotland¹⁰⁸, Rebecca Boyles¹⁰⁸, Sridhar Rathinam¹⁰⁸, Domenic Marrone¹⁰⁹, Sean Dulloo^{108,109}, Dean A. Fennell^{108,109}, Sarah Danson^{110,111}, Elaine Smith¹¹², Eustace Fontaine¹¹², Felice Granato¹¹², Juliette Novasio¹¹², Kendadai Rammohan¹¹², Leena Joseph¹¹², Paul Bishop¹¹², Rajesh Shah¹¹², Vijay Joshi¹¹² & Philip Crosbie^{46,112,113}

¹⁸Aberdeen Royal Infirmary NHS Grampian, Aberdeen, UK. ¹⁹Academic Division of Thoracic Surgery, Imperial College London, London, UK. ²⁰Royal Brompton and Harefield Hospitals, part of Guy's and St Thomas' NHS Foundation Trust, London, UK. ²¹Artificial Intelligence in Medicine (AIM) Program, Mass General Brigham, Harvard Medical School, Boston, MA, USA. ²²Department of Radiation Oncology, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. ²³Radiology and Nuclear Medicine, CARIM & GROW, Maastricht University, Maastricht, The Netherlands. ²⁴Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin,

Germany. ²⁵Berlin Institute for the Foundations of Learning and Data (BIFOLD), Berlin, Germany. ²⁶Bioinformatics and Systems Biology, Method Development and Research Infrastructure, Robert Koch Institute, Berlin, Germany. ²⁷Birmingham Acute Care Research Group, Institute of Inflammation and Ageing, University of Birmingham, Birmingham, UK. ²⁸Bill Lyons Informatics Centre, University College London Cancer Institute, London, UK. ²⁹Medical Genomics, University College London Cancer Institute, London, UK. ³⁰Cancer Research Centre, University of Leicester, Leicester, UK. ³¹Cancer Research UK & UCL Cancer Trials Centre, London, UK. ³²Department of Thoracic Surgery, Respiratory Center, Toranomon Hospital, Tokyo, Japan. ³³Advanced Sequencing Facility, The Francis Crick, London, UK. ³⁴Computational Cancer Genomics Research Group, University College London Cancer Institute, London, UK. ³⁵University College London Hospitals, London, UK. ³⁶Department of Molecular Medicine, Aarhus University Hospital, Aarhus, Denmark. ³⁷Department of Clinical Medicine, Aarhus University, Aarhus, Denmark. ³⁸Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark. ³⁹Immune Regulation and Tumour Immunotherapy Group, Cancer Immunology Unit, Research Department of Haematology, University College London Cancer Institute, London, UK. ⁴⁰Cancer Research UK Scotland Institute, Glasgow, UK. ⁴¹Institute of Cancer Sciences, University of Glasgow, Glasgow, UK. ⁴²NHS Greater Glasgow and Clyde Pathology Department, Queen Elizabeth University Hospital, Glasgow, UK. ⁴³The Institute of Cancer Research, London, UK. ⁴⁴Centre for Medical Image Computing, Department of Medical Physics and Biomedical Engineering, London, UK. ⁴⁵CRUK Manchester Institute Cancer Biomarker Centre, University of Manchester, Manchester, UK. ⁴⁶Cancer Research UK Lung Cancer Centre of Excellence, University of Manchester, Manchester, UK. ⁴⁷Danish Cancer Institute, Copenhagen, Denmark. ⁴⁸Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, USA. ⁴⁹Department of Bioinformatics, Semmelweis University, Budapest, Hungary. ⁵⁰Department of Physics of Complex Systems, ELTE Eötvös Loránd University, Budapest, Hungary. ⁵¹Department of Haematology, University College London Hospitals, London, UK. ⁵²Cancer Immunology Unit, Research Department of Haematology, University College London Cancer Institute, London, UK. ⁵³Department of Medical Oncology, Aberdeen Royal Infirmary NHS Grampian, Aberdeen, UK. ⁵⁴University of Aberdeen, Aberdeen, UK. ⁵⁵Department of Medical Physics and Bioengineering, University College London Cancer Institute, London, UK. ⁵⁶The Whittington Hospital NHS Trust, London, UK. ⁵⁷Department of Oncology, University Hospital Southampton NHS Foundation Trust, Southampton, UK. ⁵⁸Department of Pathology, Aberdeen Royal Infirmary NHS Grampian, Aberdeen, UK. ⁵⁹Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. ⁶⁰Department of Radiology, University College London Hospitals, London, UK. ⁶¹UCL Respiratory, Department of Medicine, University College London, London, UK. ⁶²Department of Thoracic Surgery, University College London Hospital NHS Trust, London, UK. ⁶³Division of Cancer Sciences, The University of Manchester and The Christie NHS Foundation Trust, Manchester, UK. ⁶⁴Experimental Histopathology, The Francis Crick Institute, London, UK. ⁶⁵Garvan Institute of Medical Research, Darlinghurst, New South Wales, Australia. ⁶⁶Genomics Science Technology Platform, The Francis Crick Institute, London, UK. ⁶⁷Golden Jubilee National Hospital, Clydebank, UK. ⁶⁸Guy's and St Thomas' NHS Foundation Trust, London, UK. ⁶⁹Independent Cancer Patient's Voice, London, UK. ⁷⁰Institute for Computational Cancer Biology, Center for Integrated Oncology (CIO), Cancer Research Center Cologne Essen (CCCE), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany. ⁷¹Institute of Infection, Immunity & Inflammation, University of Glasgow, Glasgow, UK. ⁷²Institute of Nuclear Medicine, Division of Medicine, University College London, London, UK. ⁷³Institute of Structural and Molecular Biology, University College London, London, UK. ⁷⁴Integrated Radiology Department, North-Buda St. John's Central Hospital, Budapest, Hungary. ⁷⁵Institute of Nuclear Medicine, University College London Hospitals, London, UK. ⁷⁶Integrative Cancer Genomics Laboratory, VIB Center for Cancer Biology, Leuven, Belgium. ⁷⁷VIB Center for AI & Computational Biology, Leuven, Belgium. ⁷⁸Leicester Medical School, University of Leicester, Leicester, UK. ⁷⁹Liverpool Heart and Chest Hospital, Liverpool, UK. ⁸⁰Lungs for Living Research Centre, UCL Respiratory, Department of Medicine, University College London, London, UK. ⁸¹Department of Thoracic Medicine, University College London Hospitals, London, UK. ⁸²Manchester Cancer Research Centre Biobank, Manchester, UK. ⁸³Manchester University NHS Foundation Trust, Manchester, UK. ⁸⁴Experimental Oncology, Institute for Oncology and Radiology of Serbia, Belgrade, Serbia. ⁸⁵NHS Greater Glasgow and Clyde, Glasgow, UK. ⁸⁶Princess Alexandra Hospital, The Princess Alexandra Hospital NHS Trust, Harlow, UK. ⁸⁷National Heart and Lung Institute, Imperial College, London, UK. ⁸⁸Royal Free London NHS Foundation Trust, London, UK. ⁸⁹Royal Surrey Hospital, Royal Surrey Hospitals NHS Foundation Trust, Guildford, UK. ⁹⁰University of Surrey, Guildford, UK. ⁹¹School of Cancer Sciences, University of Glasgow, Glasgow, UK. ⁹²Beatson Institute for Cancer Research, University of Glasgow, Glasgow, UK. ⁹³Queen Elizabeth University Hospital, Glasgow, UK. ⁹⁴Singleton Hospital, Swansea Bay University Health Board, Swansea, UK. ⁹⁵The Christie NHS Foundation Trust, Manchester, UK. ⁹⁶The Francis Crick Institute, London, UK. ⁹⁷Department of Infectious Disease, Faculty of Medicine, Imperial College London, London, UK. ⁹⁸Department of Neurosurgery, National Hospital for Neurology and Neurosurgery, London, UK. ⁹⁹The NIHR Southampton Biomedical Research Centre, University Hospital Southampton NHS Foundation Trust, Southampton, UK. ¹⁰⁰The University of Texas MD Anderson Cancer Center, Houston, TX, USA. ¹⁰¹University College London Cancer Institute, London, UK. ¹⁰²Department I of Internal Medicine, University Hospital of Cologne, Cologne, Germany. ¹⁰³University College London Department of Epidemiology and Health Care, London, UK. ¹⁰⁴National Cancer Centre, Singapore City, Singapore. ¹⁰⁵University Hospital Birmingham NHS Foundation Trust, Birmingham, UK. ¹⁰⁶Institute of Immunology and Immunotherapy, University of Birmingham, Birmingham, UK. ¹⁰⁷University Hospital Southampton NHS Foundation Trust, Southampton, UK. ¹⁰⁸University Hospitals of Leicester NHS Trust, Leicester, UK. ¹⁰⁹University of Leicester, Leicester, UK. ¹¹⁰University of Sheffield, Sheffield, UK. ¹¹¹Sheffield Teaching Hospitals NHS Foundation Trust, Sheffield, UK. ¹¹²Wythenshawe Hospital, Manchester University NHS Foundation Trust, Wythenshawe, UK. ¹¹³Division of Infection, Immunity and Respiratory Medicine, University of Manchester, Manchester, UK.

Methods

The TRACERx421 data

The TRACERx421 samples used in the study have been described in previously published manuscripts^{14,16}. The design of the TRACERx study has been approved by an independent research ethics committee (13/LO/1546) and the ClinicalTrials.gov number is NCT01888601. Informed consent for entry into the TRACERx study was mandatory and obtained from every patient. The purity and ploidy estimates, histological subtypes, lung cancer genes and phylogenetic trees used in this study were taken from a previous TRACERx study¹⁶. Transcripts per million (TPMs) estimates were taken from a previously published TRACERx study¹⁴. The classification of primary tumors that did and did not metastasize, as well as the classification of seeding regions, was taken from a previously published TRACERx study¹⁵. Only primary tumor and non-lymph node regions with purity and ploidy estimates were used in this study. The consort diagram of the TRACERx samples used in the study is shown in Extended Data Fig. 1.

The TCGA cohort

MHC Hammer was run on lung and breast samples from the TCGA dataset^{17–19}. We implemented the following thresholds for a sample to be included in our study:

- Whole-exome sequencing (WES) samples with less than 5,000,000 paired aligned reads or an alignment rate of less than 0.8 were excluded.
- RNA sequencing (RNA-seq) samples with an alignment rate of less than 0.6 were excluded.
- Formalin-fixed paraffin-embedded (FFPE) samples and metastatic samples were excluded.
- Tumor samples without a purity and ploidy solution were excluded.
- Samples without a matched WES germline sample that passed our filters were excluded.

In the cases where a TCGA sample was sequenced multiple times, we selected a single sequencing run to use in our study. For the WES tumor samples, we prioritized choosing a non-whole-genome amplification (non-WGA) sample over a WGA sample³¹, and then prioritized the samples by the number of paired and aligned reads. For the WES germline samples, we also prioritized non-WGA over WGA, then blood samples over solid tissue samples and finally prioritized the samples by the number of paired and aligned reads. For the RNA-seq samples, we chose the sample with the highest number of paired and aligned reads.

The purity and ploidy solutions for the TCGA cohort were estimated using ASCAT³² and taken from https://github.com/VanLoo-lab/ascats/tree/master/ReleasedData/TCGA_SNP6_hg38. The methylation array data and TPM data used in this study were downloaded from Genomic Data Commons (GDC). The lung histological subtypes (LUAD and LUSC) were taken from GDC, and the breast subtypes (ER+, ER- and TNBC) were taken from cBioPortal.

The consort diagram for the TCGA cohort is shown in Extended Data Fig. 2.

The GTEx cohort

We ran MHC Hammer on the normal lung and breast RNA-seq samples from the GTEx dataset²⁰. We implemented the following thresholds for a sample to be included in our study:

- WES germline samples with less than 5,000,000 paired aligned reads or an alignment rate of less than 0.8 were excluded.
- RNA-seq samples with an alignment rate of less than 0.6 were excluded.
- Samples without a matched WES germline sample that passed our filters were excluded.

The consort diagram for the GTEx cohort is shown in Extended Data Fig. 3.

Validation of allele-specific HLA alternative splicing

To validate our HLA alternative splicing pipeline, we used allele-specific PCR amplification. We performed this for four tumor regions and one normal sample from two patients (CRUK0061_SU_N01, CRUK0061_SU_T1-R1, CRUK0061_SU_T1-R2, CRUK0733_SU_T1-R2 and CRUK0733_SU_T1-R6). RNA-seq data were available for four of these samples (CRUK0061_SU_N01, CRUK0061_SU_T1-R1, CRUK0733_SU_T1-R2 and CRUK0733_SU_T1-R6), and MHC Hammer identified exon 5 skipping in an HLA-C allele in all four samples.

To amplify each allele, we used allele-specific primers that have been described previously¹², and the fragment sizes were confirmed via agarose gel electrophoresis (Supplementary Fig. 1a). These PCR products were then cloned using a TA cloning kit (Invitrogen), where the wild-type and novel alternatively spliced transcripts were subsequently validated through Sanger sequencing (Supplementary Fig. 1b).

Neoantigen calls

Patient-specific HLA haplotype predictions were obtained using HLA-HD³³ (version 1.2.1). NetMCHpan4.1 (ref. 34) was run on 9–11 neopeptides derived from nonsynonymous mutations across the TRACERx421 cohort, taking into account patient-specific HLA types. A cutoff of 0.5 in the eluted ligand rank was applied to define whether a peptide is bound to a specific HLA type. An observed nonsynonymous mutation is deemed a neoantigen binding to a specific HLA if at least one of its neopeptides is considered a binder.

Danaher estimates of CD8⁺ T cell infiltration

The amount of CD8 T cell infiltration was estimated using the Danaher method²⁷. To do this, TPM values of the *CD8A* and *CD8B* genes were first converted to log₂, and the mean log₂ value across the two genes was taken for each sample.

Calling alternative splicing in all protein-coding genes

To call alternative splicing in all protein-coding genes, we used the STAR aligner with a two-pass alignment^{35,36} and the GRCh38 reference sequence to generate a set of splice junctions in the TRACERx samples. Novel splice junctions were defined as those not present in the GRCh38 RefSeq GTF file (<https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/genes/>). To be considered as present in a tumor region, a novel splice junction required at least 20 uniquely mapping reads in any region from the tumor and at least two uniquely mapping reads in the given region. To be considered somatic, the novel splice junction could not be present in the patient-matched normal sample.

The MHC Hammer pipeline

See Supplementary Note for a detailed overview of the MHC Hammer pipeline. The following MHC Hammer parameters were used to generate the data in this study:

- The HLA reference files were created using the ImMunoGenetics (IMGT) database version 3.38 (ref. 37).
- The library size was estimated as the number of paired and aligned reads in the input BAM files (include_unmapped_reads_in_library_size = FALSE).
- The HLA FASTQ files were created by filtering the input BAM files to include all unmapped reads, reads that mapped to chromosome 6 or any alternate contig or reads that contained a 30-mer sequence from the IMGT database (unmapped_reads = TRUE, contig_reads = TRUE, fish_reads = TRUE).
- When filtering the HLA allele BAM files, reads with more than one mismatch to the patient-specific reference were removed (max_mismatch = 1).
- In the estimation of allelic copy number and DNA AIB, filtered SNPs required a read depth of at least 30 in TRACERx (min_depth = 30) or 5 in TCGA (min_depth = 5).

- After the first STAR alignment, splice junctions required at least two supporting reads to be included in the cohort of splice junctions in the second STAR alignment (`uniq_num_across_junc = 2`).

Filters implemented for HLA WES analysis. To be included in the HLA DNA analysis, including HLA copy number, AIB and LOH calls, a gene must pass the following filters:

- Have at least ten SNPs that pass the minimum read depth of 30 in the TRACERx samples or 5 in the TCGA samples.
- Both alleles of the gene must have an expected depth of at least 10. The expected depth estimates the depth of the reads that are coming solely from the cancer cells (see Supplementary Note for more details).
- The 95% confidence interval in the allelic copy number, calculated using the R function `t.test`, must be less than 2.5.

Filters implemented for somatic HLA mutations. HLA allelic mutations were classified as high-impact and damaging if the Ensembl Variant Effect Predictor (VEP)³⁸ consequence included at least one of ‘stop_gained’, ‘frameshift_variant’, ‘start_lost’ or ‘stop_lost’. To be considered in our analysis, mutations had to be classified as ‘PASS’ using the Genome Analysis ToolKit (GATK)³⁹ `FilterMutectCalls` function, have ten reads supporting the alternate allele, and fall in a sample and gene that passed the MHC Hammer WES filters.

Filters implemented for HLA RNA analysis. To be included in the HLA RNA analysis, including RNA AIB, allelic expression, allelic repression and alternative splicing, a gene must pass the DNA analysis filters and

- Have at least ten SNPs in the exon sequence.
- Have no more than 50% of reads mapping to both alleles of the same gene.
- Have no more than 5% of reads mapping to multiple HLA genes.

Filters implemented for HLA alternative splicing. We excluded from our analysis any novel splice junction detected in the first or last exons of an HLA gene. To be included in our analysis, novel splice junctions had to be classified as one of the following: complete exon skipping, partial exon skipping, partial intron retention or complete intron retention. In addition, we applied the following depth filters:

- In the GTEx cohort—to be included in our analysis, a novel splice junction required at least two uniquely mapping reads in that sample. In addition, the novel splice junction needed to be identified in another sample from the same patient with at least 20 uniquely mapping reads.
- In the TRACERx and TCGA cohorts—to be included in our analysis, a novel splice junction needed to be classified as either ‘tumor-enriched’ or ‘tumor-depleted’ and also be identified in at least one tumor region or matched normal from the patient with at least 20 uniquely mapping reads.

Statistical information

All statistical tests were performed in R (v.4.3.3). No statistical methods were used to predetermine the sample size. Tests involving comparisons of distributions were done using a two-tailed Wilcoxon test (`wilcox.test`). Tests involving the comparison of groups were done using a two-tailed Fisher’s exact test (`fisher.test`). The correlation was tested using Pearson’s correlation coefficient (`cor.test`).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The WES and RNA-seq data used during this study have been deposited at the European Genome–Phenome Archive, which is hosted by the European Bioinformatics Institute and the Centre for Genomic Regulation under the accession code [EGAS00001006494](https://www.ebi.ac.uk/ena/browser/view/EGAS00001006494). Access is controlled by the TRACERx data access committee. Details on how to apply for access are available on the linked page. The TRACERx data are available under controlled access so that patient privacy and data confidentiality are maintained while promoting and encouraging impactful scientific discovery. The data access committee aims to reply to requests within 1 week.

Code availability

Code to run the MHC Hammer pipeline can be found at <https://github.com/McGranahanLab/mhc-hammer>.

Code to recreate the manuscript figure can be found at <https://zenodo.org/records/13388455> (ref. 40).

References

- Buckley, A. R. et al. Pan-cancer analysis reveals technical artifacts in TCGA germline variant calls. *BMC Genomics* **18**, 458 (2017).
- Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
- Kawaguchi, S., Higasa, K., Shimizu, M., Yamada, R. & Matsuda, F. HLA-HD: an accurate HLA typing algorithm for next-generation sequencing data. *Hum. Mutat.* **38**, 788–797 (2017).
- Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454 (2020).
- Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- Veeneman, B. A., Shukla, S., Dhanasekaran, S. M., Chinnaiyan, A. M. & Nesvizhskii, A. I. Two-pass alignment improves novel splice junction quantification. *Bioinformatics* **32**, 43–49 (2016).
- Lefranc, M.-P. et al. IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Res.* **43**, D413–D422 (2014).
- McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
- Van der Auwera, G. A. & O’Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. 1st edn (O’Reilly Media, 2020).
- Puttick, C. MHC Hammer publication code and data. <https://zenodo.org/records/13388455> (2024).

Acknowledgements

We gratefully acknowledge the individuals who participated in the TRACERx, TCGA and GTEx studies. We are grateful for assistance from the scientific computing facilities at the Francis Crick Institute. We thank all site personnel, investigators, funders and industry partners who supported the generation of the data within this study. The results published here are in part based on data generated by the TCGA Research Network and the GTEx study. The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health and by NCI, NHGRI, NHLBI, NIDA, NIMH and NINDS. The TRACERx study (ClinicalTrials.gov registration: NCT01888601) is sponsored by University College London (UCL/12/O279) and has been approved by an independent Research Ethics Committee (13/LO/1546). TRACERx is funded by Cancer Research UK (C11496/A17786) and coordinated through the Cancer Research UK and UCL Cancer Trials Centre, which has a core grant from CRUK (C444/A15953). This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research

UK (CC2041), the UK Medical Research Council (CC2041) and the Wellcome Trust (CC2041). This work was also supported by the Cancer Research UK Lung Cancer Centre of Excellence and the CRUK City of London Centre Award (C7893/A26233), as well as the UCL Experimental Cancer Medicine Centre. C.P. is supported by the Breast Cancer Research Foundation (BCRF 23-157). M.M.L. is supported by the Rosetrees Trust. F.G.C. is supported by the Cancer Research UK Brain Tumor Award (A28592). C.M.-R. is supported by the Rosetrees Trust (M630) and the Wellcome Trust. K.K.D. is supported by funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie (grant agreement 101024529) and supported by a ZonMW Rubicon fellowship (20-45200-98-20102). N.K. is supported by the Breast Cancer Research Foundation (BCRF 23-157), the Rosetrees Trust and Cancer Research UK. K.L. is funded by the UK Medical Research Council (MR/V033077/1), the Rosetrees Trust and Cotswold Trust (A2437), the Royal Marsden Cancer Charity (thanks to the R. Russell family and Macfarlanes donations) and the Melanoma Research Alliance and Cancer Research UK (C69256/A30194). P.V.L. is supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (CC2008), the UK Medical Research Council (CC2008) and the Wellcome Trust (CC2008); is a CPRIT Scholar in Cancer Research and acknowledges CPRIT grant support (RR210006). M.J.-H. is a CRUK Career Establishment Awardee and has received funding from CRUK, IASLC International Lung Cancer Foundation, Lung Cancer Research Foundation, Rosetrees Trust, UKI NETs, National Institute for Health Research (NIHR) and NIHR UCLH Biomedical Research Centre. S.A.Q. is funded by a CRUK Senior Cancer Research Fellowship (C36463/A22246), CRUK Biotherapeutic Program grant (C36463/A20764) and Cancer Research UK Brain Tumor Award (A28592). C.S. is a Royal Society Napier Research Professor (RSRP\R\210001). His work is supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (CC2041), the UK Medical Research Council (CC2041) and the Wellcome Trust (CC2041). For Open Access, the author has applied a CC BY public copyright license to any author-accepted manuscript version arising from this submission. C.S. is funded by Cancer Research UK (TRACERx (C11496/A17786), PEACE (C416/A21999) and CRUK Cancer Immunotherapy Catalyst Network); Cancer Research UK Lung Cancer Centre of Excellence (C11496/A30025); the Rosetrees Trust, Butterfield and Stoneygate Trusts; Novo Nordisk Foundation (ID16584); Royal Society Professorship Enhancement Award (RP/EA/180007 and RF\ERE\231118)); NIHR University College London Hospitals Biomedical Research Centre; the Cancer Research UK-University College London Centre; Experimental Cancer Medicine Centre; the Breast Cancer Research Foundation (US; BCRF-22-157); Cancer Research UK Early Detection and Diagnosis Primer Award (grant EDDPMA-Nov21/100034); the Mark Foundation for Cancer Research Aspire Award (grant 21-029-ASP) and ASPIRE Phase II award (grant 23-034-ASP). C.S. is in receipt of an ERC Advanced Grant (PROTEUS) from the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement 835297). N.M. is a Sir Henry Dale Fellow, jointly funded by the Wellcome Trust and the Royal Society (grant 211179/Z/18/Z) and also receives funding from Cancer Research UK, Rosetrees and the NIHR BRC at University College London Hospitals and the CRUK University College London Experimental Cancer Medicine Centre.

Author contributions

C.P. performed pipeline development, analysis and wrote the manuscript. T.P.J. performed pipeline development. M.M.L. performed analysis and wrote the manuscript. F.G.C., J.L., M.V.G. and A.R. designed and performed wet lab validation. O.P., C.M.-R., R.B., K.K.D., J.R.M.B. and R.R. provided bioinformatic support. N.K., K.L., R.S., D.A.M., P.V.L., M.J.-H. and S.A.Q. supervised the study. C.S. and N.M. supervised the study and wrote the manuscript.

Competing interests

C.P. holds a patent pending in determining HLA disruption (PCT/EP2023/059039). K.K.D. provided consultancy services to Achilles Therapeutics UK. N.K. receives research support from AstraZeneca. K.L. has a patent on InDel burden and CPI response pending and speaker fees from Roche tissue diagnostics, research funding from CRUK TDL-Ono-LifeArc alliance, Genesis Therapeutics and consulting roles with Ellipses Pharma, Monopteros and Kynos Therapeutics. D.A.M. reports speaker fees from Eli Lilly, AstraZeneca and Takeda Pharmaceuticals; consultancy fees from AstraZeneca, Thermo Fisher Scientific, Takeda Pharmaceuticals, Amgen, Janssen, MIM Software, Bristol Myers Squibb and Eli Lilly; and has received educational support from Takeda Pharmaceuticals and Amgen. M.J.-H. has received funding from CRUK, NIH National Cancer Institute, IASLC International Lung Cancer Foundation, Lung Cancer Research Foundation, Rosetrees Trust, UKI NETs and NIHR. M.J.-H. has consulted for, and is a member of, the Achilles Therapeutics Scientific Advisory Board and Steering Committee and has received speaker honoraria from Pfizer, Astex Pharmaceuticals, Oslo Cancer Cluster, Bristol Myers Squibb and Genentech. M.J.-H. is listed as a co-inventor on a European patent application relating to methods to detect lung cancer PCT/US2017/028013). This patent has been licensed to commercial entities, and under terms of employment, M.J.-H. is due a share of any revenue generated from such license(s) and is also listed as a co-inventor on the GB priority patent application (GB2400424.4) with title—Treatment and Prevention of Lung Cancer. C.S. acknowledges grants from AstraZeneca, Boehringer-Ingelheim, Bristol Myers Squibb, Pfizer, Roche-Ventana, Invitae (previously Archer Dx—collaboration in minimal residual disease sequencing technologies), Ono Pharmaceutical and Personalis. He is the chief investigator for the AZ MeRmaid 1 and 2 clinical trials and is the Steering Committee Chair. He is also the cochief investigator of the NHS Galleri trial funded by GRAIL and a paid member of GRAIL's Scientific Advisory Board. He receives consultant fees from Achilles Therapeutics (also a SAB member), Bicycle Therapeutics (also a SAB member), Genentech, Medixis, China Innovation Centre of Roche (CICoR) formerly Roche Innovation Centre—Shanghai, Metabomed (until July 2022), Relay Therapeutics SAB member, Saga Diagnostics SAB member and the Sarah Cannon Research Institute. He has received honoraria from Amgen, AstraZeneca, Bristol Myers Squibb, GlaxoSmithKline, Illumina, MSD, Novartis, Pfizer and Roche-Ventana. He has previously held stock options in Apogen Biotechnologies and GRAIL; currently has stock options in Epic Bioscience, Bicycle Therapeutics, and Relay Therapeutics; and has stock options and is cofounder of Achilles Therapeutics. He declares a patent application for methods to lung cancer (PCT/US2017/028013); targeting neoantigens (PCT/EP2016/059401); identifying patent response to immune checkpoint blockade (PCT/EP2016/071471); methods for lung cancer detection (US20190106751A1); identifying patients who respond to cancer treatment (PCT/GB2018/051912); determining HLA LOH (PCT/GB2018/052004); predicting survival rates of patients with cancer (PCT/GB2020/050221); and methods and systems for tumor monitoring (PCT/EP2022/077987). He is an inventor of a European patent application (PCT/GB2017/053289) relating to assay technology to detect tumor recurrence. This patent has been licensed to a commercial entity, and under their terms of employment, he is due a revenue share of any revenue generated from such license(s). N.M. has stock options in and has consulted for Achilles Therapeutics and holds a European patent in determining HLA LOH (PCT/GB2018/052004), a patent pending in determining HLA disruption (PCT/EP2023/059039), and is a co-inventor to a patent to identify responders to cancer treatment (PCT/GB2018/051912). The remaining authors declare no competing interests.

Additional information

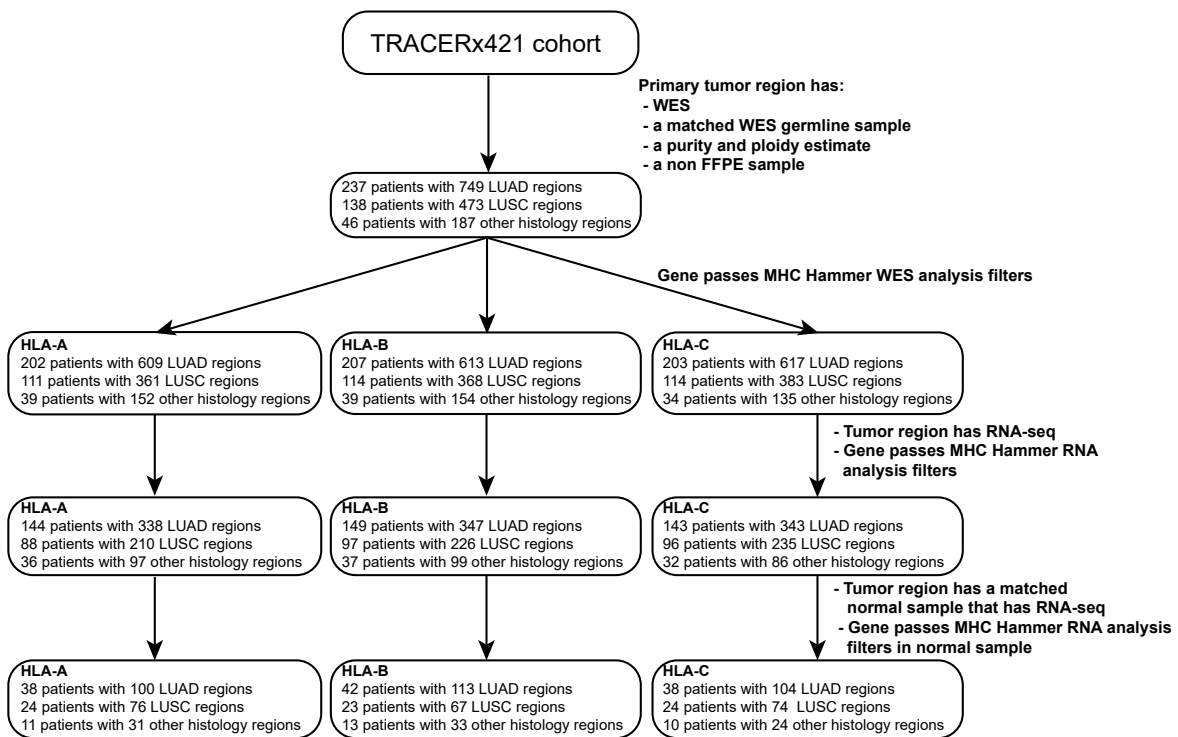
Extended data is available for this paper at <https://doi.org/10.1038/s41588-024-01883-8>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41588-024-01883-8>.

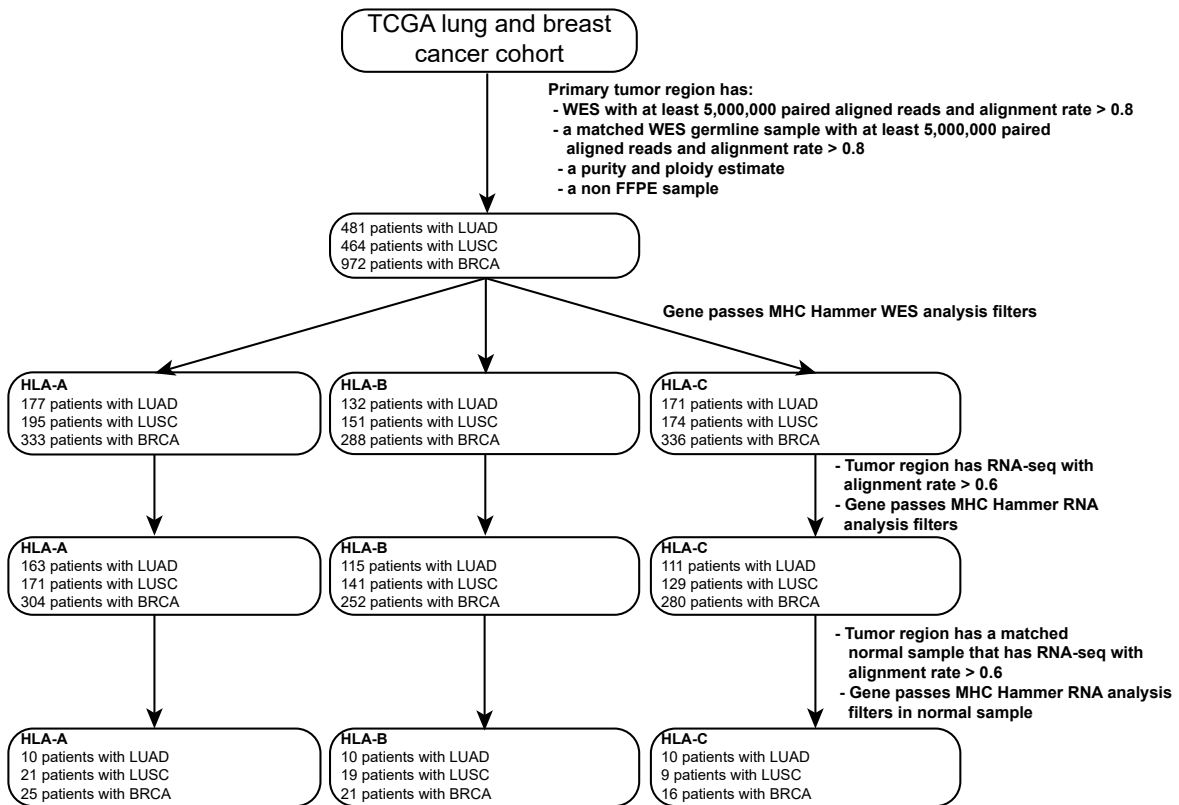
Correspondence and requests for materials should be addressed to Charles Swanton or Nicholas McGranahan.

Peer review information *Nature Genetics* thanks Justin Gainor and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

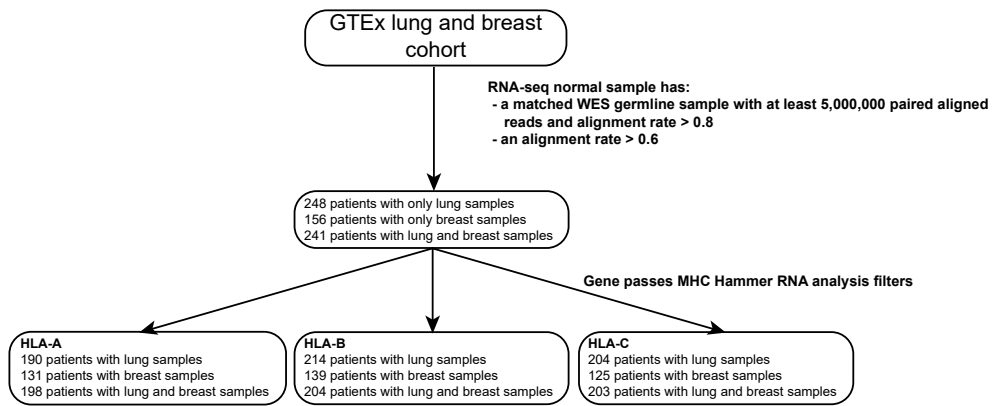


Extended Data Fig. 1 | The TRACERx421 cohort. Consort diagram outlining the TRACERx421 samples used in this study. The MHC Hammer WES and RNA-seq analysis filters are outlined in the Methods. FFPE, formalin-fixed paraffin-embedded; WES, whole-exome sequencing; RNA-seq, RNA sequencing; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

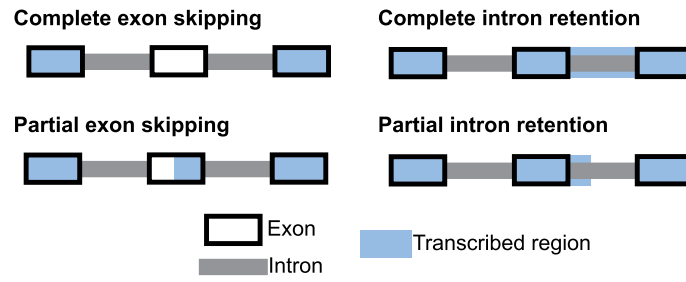


Extended Data Fig. 2 | The TCGA lung and breast cancer cohort. Consort diagram outlining the TCGA lung and breast samples used in this study. The MHC Hammer WES and RNA-seq analysis filters are outlined in the Methods.

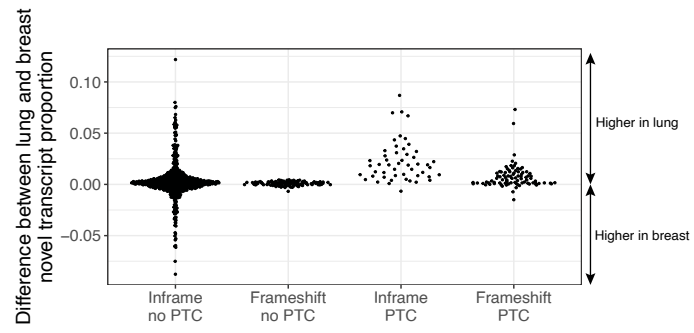
FFPE: formalin-fixed paraffin-embedded; WES: whole-exome sequencing; RNA-seq: RNA sequencing; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; BRCA, breast cancer.



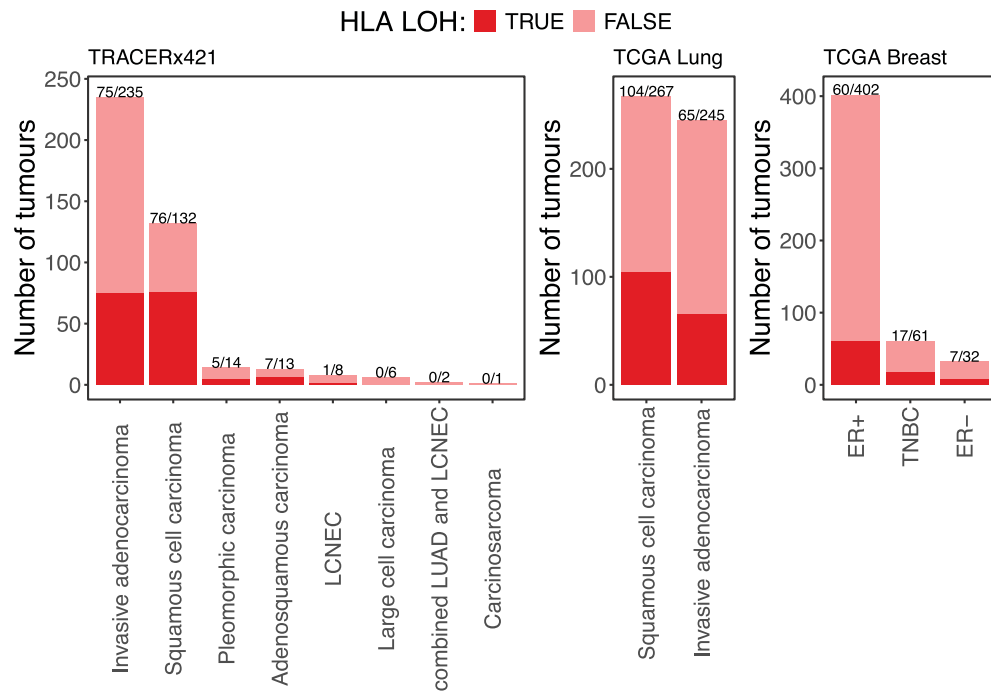
Extended Data Fig. 3 | The GTEx cohort. Consort diagram outlining the GTEx lung and breast samples used in this study. The MHC Hammer RNA-seq analysis filters are outlined in the Methods. WES: whole-exome sequencing; RNA-seq: RNA sequencing.



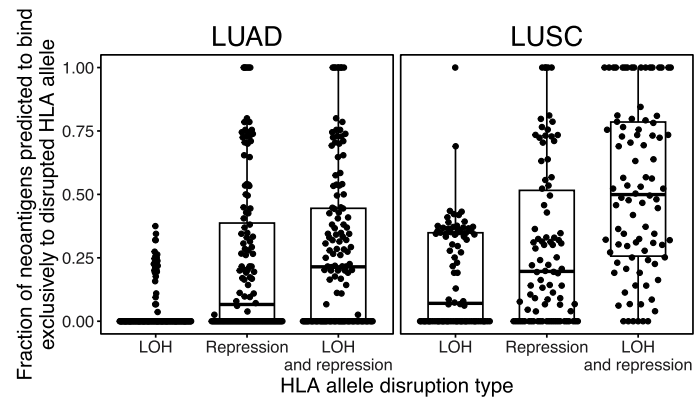
Extended Data Fig. 4 | Types of alternative splicing detected by MHC Hammer. MHC Hammer will detect 4 different types of alternative splicing: complete exon skipping, complete intron retention, partial exon skipping and partial intron retention.



Extended Data Fig. 5 | The difference between the lung and breast novel transcript proportion. For alternative splicing events that were found in both the lung and breast tissue of the same patient, the difference in the novel transcript proportion between the two tissues is shown.



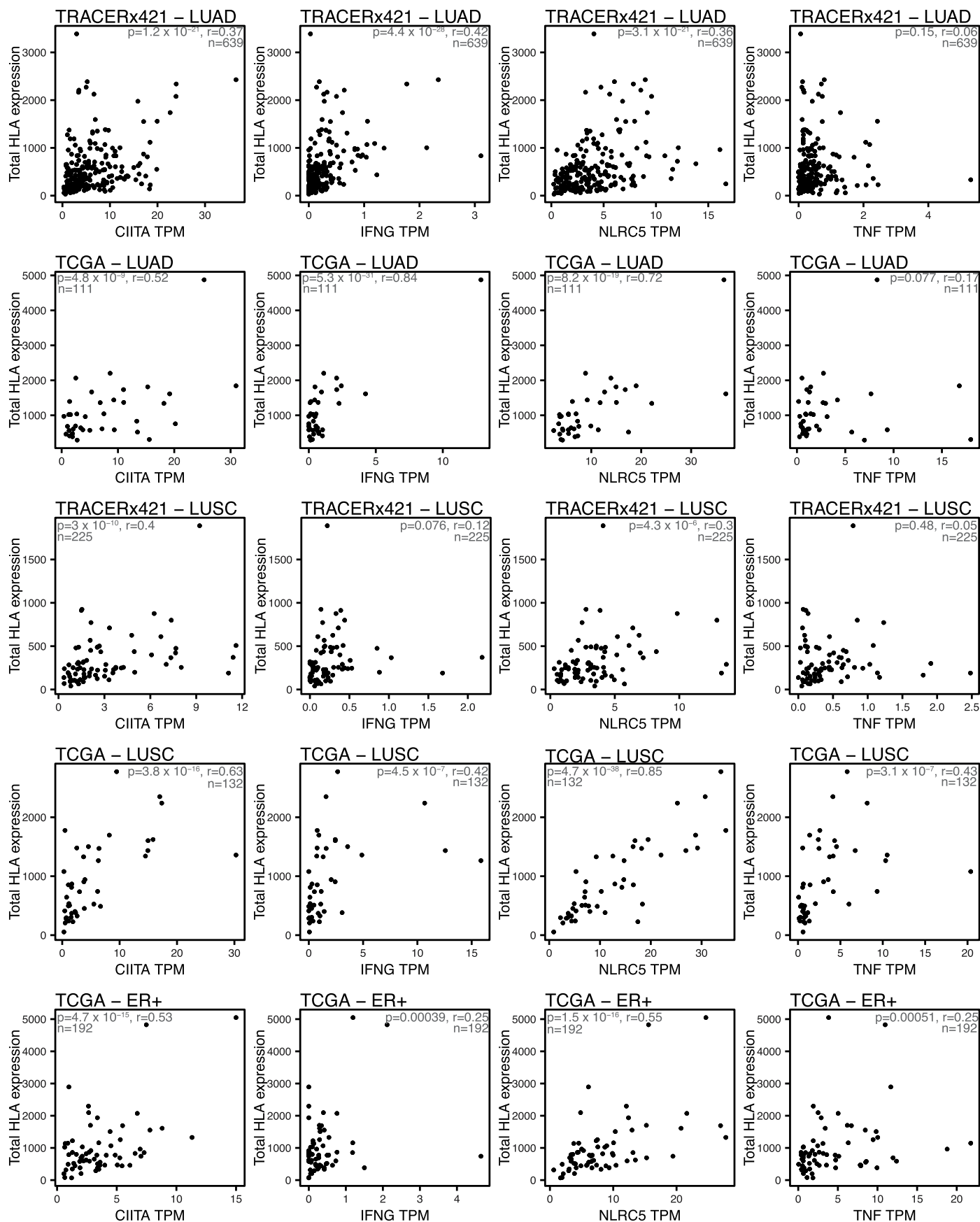
Extended Data Fig. 6 | HLA LOH rates in lung and breast cancer. The rate of HLA LOH in the TRACERx421 lung cohort and the TCGA lung and breast cancer cohorts. LOH: loss of heterozygosity; LUAD: lung adenocarcinoma; LCNEC: large cell neuroendocrine carcinoma; ER+: estrogen receptor positive; ER-: estrogen receptor negative; TNBC: triple-negative breast cancer.



Extended Data Fig. 7 | Neoantigen silencing due to HLA LOH and repression.

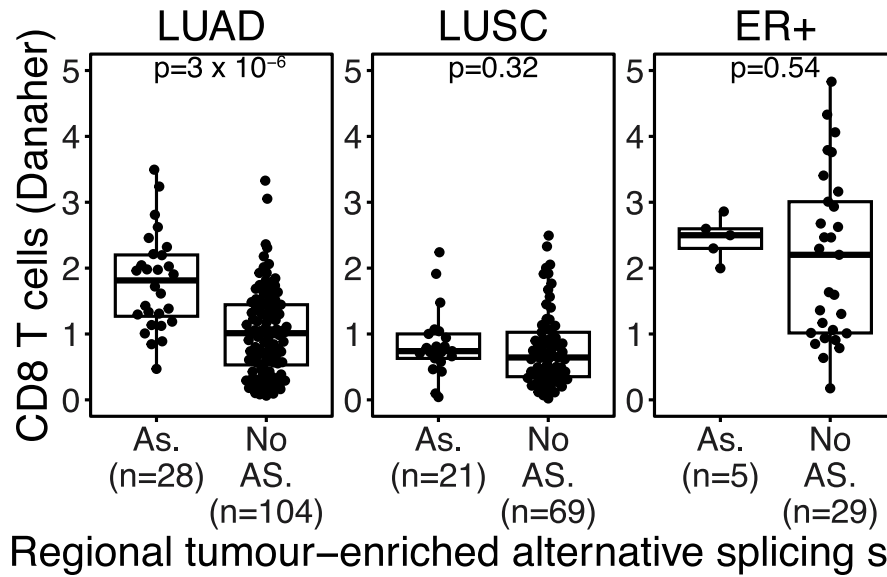
For each tumor region, the fraction of putative neoantigens predicted to bind exclusively to HLA alleles subject to LOH and/or repression. Boxplot shows

median and first and third quartiles, and whiskers extend up to 1.5× IQR above and below the IQR. LOH: loss of heterozygosity; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma.



Extended Data Fig. 8 | The HLA gene regulators and HLA expression. The relationship between total HLA expression, measured by MHC Hammer, and the expression of *CIITA*, *IFNG*, *NLR5* and *TNF α* (*TNF*) in tumor samples without HLA genomic disruption. The P value and correlation coefficient (r) are

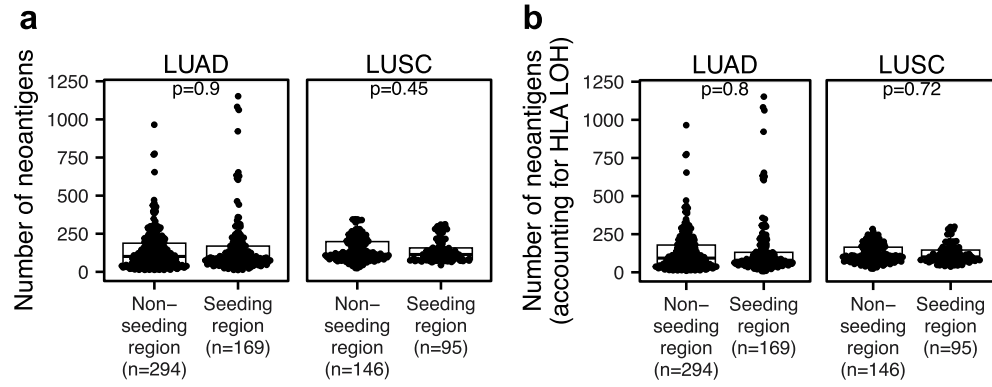
calculated using Pearson's method. TPM, transcripts per million; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; ER+, estrogen receptor positive.



Regional tumour-enriched alternative splicing status

Extended Data Fig. 9 | CD8 T cells and HLA alternative splicing. The relationship between the presence of tumor-enriched alternative splicing (AS) and the amount of CD8 T cell infiltration. P values were derived from a two-sided Wilcoxon test. Boxplot shows median and first and third quartiles, and whiskers

extend up to 1.5× IQR above and below the IQR. LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; ER+: estrogen receptor positive; AS: alternative splicing.



Extended Data Fig. 10 | Tumor neoantigen burden and metastasis seeding. **a**, The neoantigen count of primary tumor regions, split by whether they did or did not seed a metastasis. **b**, The neoantigen count of primary tumor regions, restricted to reflect peptides binding only to HLA alleles without HLA LOH, split by whether the primary tumor regions did or did not seed a metastasis. P values

were derived from a two-sided Wilcoxon test. Boxplot shows median and first and third quartiles, and whiskers extend up to $1.5 \times$ IQR above and below the IQR. LOH: loss of heterozygosity; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	No software was used to collect data
Data analysis	Code to run the MHC Hammer pipeline can be found here: https://github.com/McGranahanLab/mhc-hammer R (v4.3.3) NetMHpan (v4.1) HLA-HD (v1.7.0) samtools (v1.15.1) NovoAlign (v3.09.4) STAR (v2.7.10a) All code to recreate the figures is publicly available

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

The WES and RNAseq data used during this study has been deposited at the European Genome–phenome Archive (EGA), which is hosted by The European Bioinformatics Institute (EBI) and the Centre for Genomic Regulation (CRG) under the accession codes EGAS00001006494; access is controlled by the TRACERx data access committee. Details on how to apply for access are available on the linked page.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	<input type="text" value="Not reported on in manuscript"/>
Reporting on race, ethnicity, or other socially relevant groupings	<input type="text" value="Not reported on in manuscript"/>
Population characteristics	<input and="" cancer="" evolution="" impact="" in="" lung="" of="" selection="" subclonal="" the="" tracerx""="" type="text" value="Described in the original TRACERx manuscript - "/>
Recruitment	<input and="" cancer="" evolution="" impact="" in="" lung="" of="" selection="" subclonal="" the="" tracerx""="" type="text" value="Described in the original TRACERx manuscript - "/>
Ethics oversight	<input and="" cancer="" evolution="" impact="" in="" lung="" of="" selection="" subclonal="" the="" tracerx""="" type="text" value="Described in the original TRACERx manuscript - "/>

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	<input type="text" value="No statistical methods were used to predetermine sample size"/>
Data exclusions	<input type="text" value="Some data were excluded based on pre-established criteria, which is described in the methods section."/>
Replication	<input type="text" value="The data in this study is not the result of an experimental set up, but is based on the available samples in the TRACERx, TCGA and GTEx datasets."/>
Randomization	<input type="text" value="Not applicable as this is based on observational data."/>
Blinding	<input type="text" value="Not applicable as this is based on observational data."/>

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

Methods

- n/a Involved in the study
- Antibodies
- Eukaryotic cell lines
- Palaeontology and archaeology
- Animals and other organisms
- Clinical data
- Dual use research of concern
- Plants

- n/a Involved in the study
- ChIP-seq
- Flow cytometry
- MRI-based neuroimaging

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration	NCT01888601
Study protocol	Described in the original TRACERx manuscript - "The evolution of lung cancer and impact of subclonal selection in TRACERx"
Data collection	Described in the original TRACERx manuscript - "The evolution of lung cancer and impact of subclonal selection in TRACERx"
Outcomes	Described in the original TRACERx manuscript - "The evolution of lung cancer and impact of subclonal selection in TRACERx"

Plants

Seed stocks	<i>Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.</i>
Novel plant genotypes	<i>Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.</i>
Authentication	<i>Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.</i>